Enabling Open Science publishing for Research Infrastructures via OpenAIRE: The EPOS use-case

Paolo Manghi
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
paolo.manghi@isti.cnr.it

Michele Manunta
Institute for the Electromagnetic Sensing of the Environment
National Research Council
Naples, Italy
manunta.m@irea.cnr.it

Miriam Baglioni
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
miriam.baglioni@isti.cnr.it

Alessia Bardi
Institute of Information Science and Technologies
National Research Council
Pisa, Italy
alessia.bardi@isti.cnr.it

Francesco Casu
Institute for the Electromagnetic Sensing of the Environment
National Research Council
Naples, Italy
casu.f@irea.cnr.it

Claudio De Luca
Institute for the Electromagnetic Sensing of the Environment
National Research Council
Naples, Italy
deluca.c@irea.cnr.it

Argiro Kokogiannaki
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
argirok@di.uoa.gr

***Abstract*—Research infrastructures support scientists of a research community with the tools and services they require to perform scientific activities in the digital era. They are "system of systems", often mature enough to support reproducibility of science by keeping forms of research objects. Still, often, such research objects are not published as scientific products. Publishing is limited to articles, possibly associated to datasets, and rarely to software. In this article, we present the solution proposed by the OpenAIRE infrastructure to solve this issue, with minimal efforts for research infrastructures and without renouncing any of their services or current practices. OpenAIRE offers a Research Community Dashboard service, designed to flank research infrastructures with scholarly communication services for discovery and publishing of an interlinked graph of articles, datasets, software, and other products, including research objects when these can be built. We present the Research Community Dashboard and its application in a real-case scenario, the one of the European Plate Observing System (EPOS) research infrastructure.**

Keywords—Open Science, EPOS, OpenAIRE, publishing, scholarly communication, research objects, reproducibility

I. Introduction

Research infrastructures (RIs) are focused on supporting scientists of a research community with tools and services they require to perform scientific activities in the digital era. They are "system of systems", concentrated on service provision, economy of scale, technology standards, best practices, etc., in many cases mature enough to support reproducibility of science by keeping forms of research objects. Their services, for example, can keep encodings of experiments, track provenance, and allow scientists to recover elements of scientific processes to individually re-use them or reproduce the experiment. Still, often, the idea of scientific publishing they support does not match such maturity and the concept of Open Science publishing they could easily deliver. The scientific products they publish are still mainly articles, in some cases associated to the relative datasets, and the publishing process is still performed manually, by the scientists, with very limited support to reproducibility. Disadvantages are known [1]: effective assessment of science is impeded, scientific reward is not omni-comprehensive, cost of science increases, etc.

In this article, we present the activities carried out in the OpenAIRE infrastructure [10] to support research infrastructures at overcoming their current limits with minimal efforts and without renouncing any of their services or current practices. OpenAIRE offers a Research Community Dashboard (RCD) service, designed to flank research infrastructures with scholarly communication services for discovery and publishing of an interlinked graph of articles, datasets, software, and other products, including research objects when these can be built. Scientists in a RI can delegate their RCD to aggregate all scientific products pertaining their RI, classifying them as one the aforementioned types, and identify semantic links between such products. Most importantly, the RI services can interact with the RCD to deposit scientific products, hence publish them as scholarly communication first-citizens, with attribution/citation metadata and a DOI. In particular we present the Research Community Dashboard and its application in a real-case scenario, the one of the European Plate Observing System (EPOS) research infrastructure, highlighting the effort required for the integration of the EPOS service EPOSAR in an Open Science publishing workflow. Such piloting case is being designed and implemented as part of the OpenAIRE-Advance project activities.

II. OpenAIRE and Open Science support for research communities

OpenAIRE is the European infrastructure in support of Open Science. It fosters and monitors the adoption of Open Science across Europe and beyond, at the National and international level and at the research community level. It advocates the importance and the uptake of Open Science-oriented research life-cycles and publishing workflows, in support of reproducible science, transparent assessment, and omni-comprehensive scientific reward. To this aim OpenAIRE leverages the required cultural shift via a pervasive network of people in Europe (NOADs - National Open Access Desks) and beyond ("global alignment" via CORE), and facilitates the technological shift by providing technical services and interoperability guidelines. The aggregation and information inference (text-mining) services are focused on the creation of an information graph interlinking metadata about literature, datasets, software, so-called "other products" of science, projects, funders, institutions, and scholarly communication data sources (Fig. 1) [9].

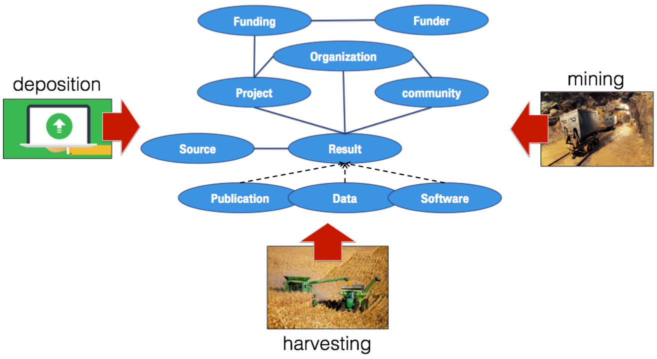
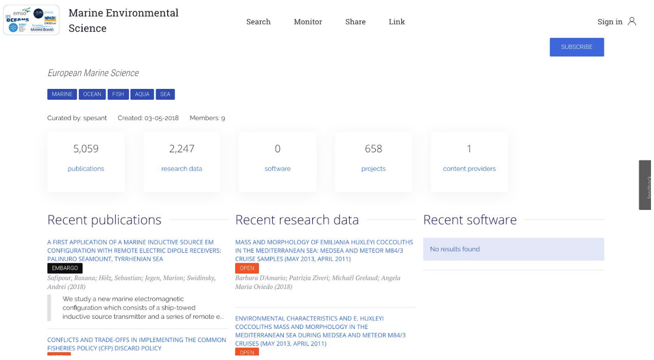


Fig 1. OpenAIRE data model and graph population modes

OpenAIRE delivers different kinds of dashboards that offer functionality for viewing, accessing and managing the OpenAIRE information graph from the perspectives of different actors: researchers, project coordinators, funder officers, institution managers, and data source managers. Among its dashboards, OpenAIRE offers the Research Community Dashboard (RCD), designed to allow scientists of any research community working on top of a research infrastructure (but also long-tail communities that do not have a dedicated RI or use multiple RIs) to keep on using their tools and services, based on their workflows and best practices, while transparently taking advantage of a common portal, offering Open Science-oriented publishing tools. The RCD is a web application (see Fig. 2) that provides researchers of a community the functionality to publish, aggregate, monitor and discover their research outputs in the OpenAIRE information graph, in order to maintain a fully-fledged view of a specific scholarly discipline. The RCD interacts with the information graph by adding a community tag to the objects of the graph that are associated to a community explicitly by users, criteria, or inference process (note that an object may belong to more than one community). In particular, the RCD serves four kinds of users: the anonymous users, the community managers, the authorized scientists, and the users of authorized RI services:

- The *anonymous users* can access search and browse functionalities and, if authorized, can access the monitoring functionalities.
- The *community managers* are the ones in charge of authorizing access to scientists willing to join the community; they can also select and configure the criteria based on which products in the OpenAIRE graph can be included in the community graph. Such criteria include: products linked to research projects associated to the community, products deposited in data sources associated to the community (e.g. thematic repositories), products whose subjects belong to community vocabularies (e.g. Mesh, DEWEY, ACM, etc), and several options of "context-propagation". Context propagation is the logic based on which if a product is associated to the community, then other products linked to it with a given semantics belong to the community too. For example: a publication linked to the community is associated with a "supplementedBy" relationship to a set of datasets; then the community "context" can also be associated to the datasets.
- The *authorized scientists* can enrich the community information graph by: adding links between products (e.g. product-project, product-product), claiming products (i.e. adding existing products to the dashboard via DOI), or by depositing a new product in Zenodo.org, the catch-all repository developed at CERN to support researchers with no repository of reference (i.e. providing the type of product, citation metadata, and relative files, in order to get a new DOI for the product).
- The *authorized services* are services of the RI that can perform, via APIs, all the actions performed by authorized scientists.



Marine Environmental Science

European Marine Science

Curated by: spesant Created: 03-05-2018 Members: 9

5,059	2,247	0	658	1
publications	research data	software	projects	content providers

Recent publications

A FIRST APPLICATION OF A MARINE INDUCTIVE SOURCE EM CONFIGURATION WITH REMOTE ELECTRIC DIPOLE RECEIVERS: PALINURO SEAMOUNT, TYRRHENIAN SEA

EMBARGO

Safipour, Roxana; Hölz, Sebastian; Jegen, Marion; Swidinsky, Andrei (2018)

We study a new marine electromagnetic configuration which consists of a ship-towed inductive source transmitter and a series of remote e...

CONFLICTS AND TRADE-OFFS IN IMPLEMENTING THE COMMON FISHERIES POLICY (CFP) DISCARD POLICY

Recent research data

MASS AND MORPHOLOGY OF EMILIANIA HUXLEYI COCCOLITHS IN THE MEDITERRANEAN SEA: MEDSEA AND METEOR M84/3 CRUISE SAMPLES (MAY 2013, APRIL 2011)

OPEN

Barbara D'Amario; Patrizia Ziveri; Michaël Grelaud; Angela Maria Oviedo (2018)

ENVIRONMENTAL CHARACTERISTICS AND E. HUXLEYI COCCOLITHS MASS AND MORPHOLOGY IN THE MEDITERRANEAN SEA DURING MEDSEA AND METEOR M84/3 CRUISES (MAY 2013, APRIL 2011)

OPEN

Recent software

No results found

Fig. 2. Snapshot of the OpenAIRE Research Community Dashboard: discovery of research products

The integration of an RI with a dedicated RCD can take place by simply configuring the RCD to match community needs and recommending RI scientists the use of the RCD to "publish the unpublished", hence manually share/publish their products in Zenodo.org or interlink their products via the RCD GUIs. Alternatively, integration can also occur at the service level, realizing a form of "on-demand publishing": RI services, on request of the scientists, transparently publish via the RCD APIs all the products scientists produce while performing their usual experiments.

III. The European Plate Observing System and Open Science Inception

The European Plate Observing System (EPOS) is a pan-European distributed Research Infrastructure for solid Earth science to support a safe and sustainable society. Through the integration of National research infrastructures and data, EPOS will allow scientists to make a step change in developing new geo-hazards and geo-resources concepts and Earth science applications to help address key societal challenges. The EPOS RIs are grouped in Thematic Core Services (TCS) that represent specific communities, such as seismology, volcanic observatories, near fault observatories, GNSS and satellite data. EPOS Integrated Core Service (ICS) will provide harmonized and integrated access to data, products, services and software made available by each TCS, in order to allow users to easily access facilities belonging to different domains of solid Earth science.

CNR-IREA is an Italian service provider of EPOS whose portfolio includes satellite Earth Observation services aimed at generating value-added products for Solid Earth applications and natural disaster analysis, prevention and mitigation. In particular, EPOS scientists can benefit from the on-demand EPOSAR service that implements an advanced Differential Synthetic Aperture Radar (SAR) Interferometry (DInSAR) technique, referred to as Parallel Small Baseline Subset (P-SBAS) approach [2][3], to detect Earth surface displacements with sub-centimetre accuracy [4]. The P-SBAS approach allows the generation of surface deformation time series and velocity maps by processing in a parallel and efficient way SAR dataset consisting of tens or hundreds of images acquired over the same investigated area [5][6]. EPOSAR outputs are extremely effective to investigate natural (earthquakes, volcanic unrests, landslides) and/or man-made (tunnelling excavations, aquifer exploitation, oil and gas storage and extraction, infrastructures monitoring) hazards. The geo-scientist exploits the displacement information generated by the EPOSAR service to investigate the trigger factors of the phenomenon; accordingly surface displacements are largely used, for example, to model the source mechanism of an earthquake, the magmatic chamber of a volcano or the structural behaviour of a dam.

EPOSAR can be accessed via the ESA's *Geohazards Exploitation Platform* (GEP) [7]. GEP is a platform that transparently offers scalable and parallel processing/analysis tools over pre-loaded satellite big datasets. Scientists can integrate their *applications,* i.e. analysis algorithms/chains like P-SBAS above for EPOSAR, and make them available as-a-service through the GEP geo-browser (Fig. 3) to other scientists.

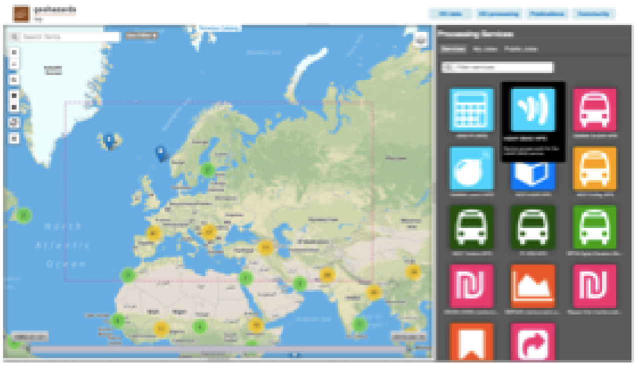


Fig. 3. EPOSAR GUI offered by GEP platform. The GUI is made up of a geobrowser, where data and products can be easily discovered through geospatial filters, and the space of available services, where the users find the applications (web-tool) to process the satellite data.

Once the scientist has selected an application/service (e.g. EPOSAR), the geo-browser supports a generic workflow (see Fig. 4) of (i) satellite image selection via a geo-spatial browser, (ii) execution of the application, (iii) visualization of results. For example, scientists can transparently select and efficiently process images of the old ESA missions (ERS-1/2 and ASAR-ENVISAT), third party data (Cosmo-Skymed and Terrasar-X) and, above all, the Copernicus Programme repository. This latter, in particular, makes available the Sentinel-1 data acquired from 2014 over the main worldwide tectonic and volcanic areas with an interferometric revisit time of 6 days. Transparently to the user, GEP executes the application workflow as a parallel job on a public/private cloud [8], by properly allocating the computing and storage resources. As shown in Fig. 4, experiments (i.e. the relative job as executed with given parameters) and results of analysis are stored into dedicated database and made available for inspection via GUI.

Specifically, the EPOSAR service workflow allows users to process the selected interferometric SAR images in unsupervised manner, and execute P-SBAS experiments by adding the specific input parameters: area of interest, reference point, and version of Digital Elevation Model (1 or 3 arcsecond SRTM).

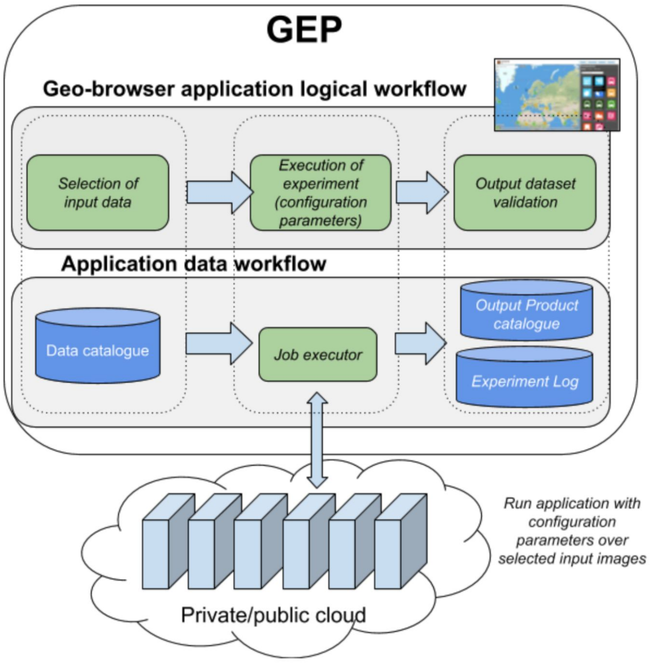

GEP
Geo-browser application logical workflow
Selection of input data
Execution of experiment (configuration parameters)
Output dataset validation
Application data workflow
Data catalogue
Job executor
Output Product catalogue
Experiment Log
Run application with configuration parameters over selected input images
Private/public cloud

Fig. 4. Generic experimental workflows supported by GEP

The P-SBAS processing chain starts copying the selected SAR images, from the archives into the virtual machines dedicated to the data processing, and retrieving ancillary information (i.e., electronic radar parameters, satellite orbital information, DEM of the area of interest). Once data unpacking has been performed, the interferometric steps (i.e. SAR image co-registration, interferogram generation, phase unwrapping operation, SVD inversion, and conversion in a geographical reference system) generate as output the deformation time-series and other ancillary information [3]. In particular, the workflow generates the following outputs: geographical coordinates, temporal coherence, mean deformation velocity, residual topography, and deformation time-series for each identified measure point. The results are provided in an ASCII file with the corresponding metadata information (ISO 19119 standard is applied); moreover some kml/kmz files are also created and provided for quick visualization.

A. *Open Science publishing with EPOSAR in EPOS*

In collaboration with OpenAIRE, CNR-IREA will integrate its EPOS services with the RCD service in order to ensure publishing of research products and experiments in a way that supports their use, reuse and reproducibility. This long term vision will be piloted in the context of the OpenAIRE-Advance project and initiated with a specific use-case, involving the EPOSAR service and its support for scientists.

Currently EPOS scientists perform experiments with EPOSAR to generate the output analysis they need to come or derive scientific conclusions. The community (beyond EPOS) has no specific best practices on how to deposit and cite the datasets resulting from EPOSAR or other EPOS services. A few journals have recently started to require the deposition of datasets in a data repository, but in general data is embedded in the article as an image or a table. This makes the overall degree of FAIR-ness and reproducibility in EPOS very low: datasets are often not shared as independent scientific products, provenance of datasets is not provided, and the experimental tools are not shared.

On the other hand, GEP preserves the history of executions, which can therefore be reproduced, and the history of the relative output time-series or velocity maps. Each experiment in EPOSAR generates and preserves the encodings required to repeat the experiment and access its result. By properly publishing such research products, the community could definitely benefit from more advanced degrees of reproducibility and transparency. The experiment product is the real added-value to this scenario, as it provides a guarantee of quality to other scientists, including scientific article reviewers: the output data accompanying an article is obtained from a recognized, repeatable, EPOS-certified workflow. Furthermore, once published with DOI and citation metadata, the experiment and dataset can be cited by others thereby increasing the scientific metrics of the authors, the visibility of the EPOSAR service, and the monitorable impact of the EPOS infrastructure.

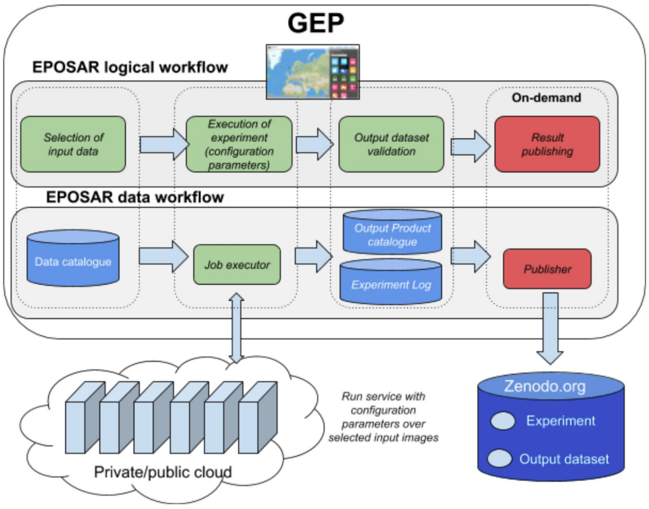


Fig. 5. EPOSAR scientific workflow integrated with OpenAIRE RCD

B. *The EPOSAR use-case in OpenAIRE*

The OpenAIRE RCD has been designed in order to take advantage of such scenarios, where the RI services are advanced but not yet focused on scholarly communication, to offer minimal-effort bridges enabling the implementation of Open Science publishing workflows. As a first step, an OpenAIRE EPOS RCD will be deployed, where scientists can identify, interlink, and claim products relevant to the EPOS community. Secondly, the EPOSAR GUI will be modified in order to let the user decide if to publish an experiment and relative results, thereby implementing a “on-demand publishing” workflow. Finally, to this aim, GEP will be equipped with a *publishing* component capable of fetching experiment and dataset information from the local databases, package them as products as required by Zenodo.org, and deposit them via APIs on behalf of the EPOSAR authorized service under the EPOS community. More specifically, the publishing component generates for each experiment to be published two research products:

- A Zenodo “other product” of type “research object” relative to the experiment, encoded as a machine readable file;
- A Zenodo “datasets” relative to the output displacement time series or the velocity maps, consisting of the files described in the previous section.

Such products will be reciprocally interlinked, have their own DOI, given citation metadata, semantics links with other products if needed, and, as a consequence of being deposited in Zenodo under the EPOS community, be discoverable/browsable through the OpenAIRE EPOS RCD. It is of course up to the users to opt when their experiment is mature enough to be published in OpenAIRE as a citable and preserved Experiment object, and eventually cite the object from any articles they may produce.

C. Technical efforts

The implementation of the EPOSAR workflow extension described above requires two interventions: a the GUI level, to enable the user to request publishing of results, and at the data flow level, to perform the deposition of the results into Zenodo.org and, indirectly, into the EPOS RCD. In the following we first describe the nature of two output research products and then describe the effort required for the EPOSAR-RCD integration.

EPOSAR for reproducible science The two products to be published require the acquisition of different kinds of metadata information. Following the Zenodo.org standard, both will be described using DataCite metadata schema[1] as recommended by the OpenAIRE guidelines for content providers.[2] Both need the minimal citation metadata, i.e. title, creators, publishing date, license, and access rights (plus other otional DataCite properties), and feature the following kinds of relationships:

- Relationship with semantics “supplementOf" to the article (if any) whose conclusions benefited from the experiment and relative output datasets;
- Relationship with semantics “refersTo” to the output dataset (if the object is the experiment) or the experiment (if the object is the output dataset) in Zenodo;
- Infrastructure concept: relationship to the EPOS infrastructure as a community in OpenAIRE (and in Zenodo); this link will associate the product to the EPOS RCD;
- Funding context: relationship to the funding grant (if any).

When focusing on the *experiment product*, the idea is to support reproducibility at the level of the service. The product payload (e.g. OGC standards, ResearchObject.org) should contain the parameters necessary to the EPOSAR service in order to repeat the experiment on request of the scientists. Such parameters include:

- The list of SAR images in the GEP data catalogue that serve as input to the experiment; images are selected via the geo-browser; according to EPOS policies, original satellite images should be preserved at their original locations (cannot be placed as products in Zenodo);
- The parameters required by the P-SBAS processing chain embedded in EPOSAR: area of interest, reference point, and version of Digital Elevation Model.

Scientist finding the product in Zenodo must be able to find a link to EPOSAR interface, which in turn, given the experiment DOI, should be able to repeat the experiment - this level of reproducibility is not platform-agnostic; in future implementations, other forms of reproducibility, based on publishing the docker of the application, will be considered. To this aim, the metadata of the experiment should also contain:

- Relationship with semantics “” to the EPOSAR service for the re-execution of the experiment; the URL should lead the the page where the user can, by inserting the DOI of the experiment, run the experiment with a click;
- Relationship with semantics with link (dc:relation with DOI, semantics “cites") to published articles describing the EPOSAR service and P-SBAS algorithm, which can both serve as documentation for the reproducibility process.

When focusing on the *output dataset product*, the publisher component will either upload the CSV file corresponding to the time series or upload a ZIP file containing the list of velocity maps.

EPOSAR upgrade: technical effort In order to allow the EPOSAR GUI, on request of the user, to perform the above actions, the GUI needs to be updated to acquire the citation metadata required to store images as datasets and the experiment as a research object: indeed EPOSAR can only partially derive from the GEP databases the metadata required by Zenodo; for example it can obtain publishing dates, but it is only aware of the user’s identity and does not provide titles for the images, the complete list of creators, license, access rights, etc. On the other hand, the implementation of the publishing component is not technically complex, as it requires, once all the metadata information is collected, the construction of a Zenodo deposition package and a few calls to the Zenodo APIs, in order to deposit the objects, obtain the relative DOIs and subsequently update their reciprocal links. Zenodo’s documentation[3] provides simple and intuitive XML/JSON packaging instructions to ingest products into the repository.

IV. Conclusions

In this paper we have shown a solution proposed in the OpenAIRE infrastructure to the notion of “on-demand publishing”, the idea of delegating RI services to publish on behalf of the user to support Open Science publishing. To

[1] DataCite schema, *https://schema.datacite.org/*

[2] OpenAIRE guidelines for content providers, *http://guidelines.openaire.eu*

[3] Zenodo documentation for developers, *http://developers.zenodo.org/*

this aim OpenAIRE has designed the solution architecture in collaboration with the EPOS infrastructure, in support of the experimental workflows possible with the EPOSAR service. The next months will be dedicated to the implementation of the integration, to be completed in 2019. Interestingly, the implementation of the publishing component will be indirectly serving all applications integrated into GEP as offered as-a-service. As a consequence, we are optimistic but pragmatic in saying that all GEP-supported services, made available via the GEP geo-browser, will benefit from the same on-demand Open Science publishing workflow.

Acknowledgements This work is partly funded by the OpenAIRE-Advance H2020 project (grant number: 777541; call: H2020-EINFRA-2017), the OpenAIRE-Connect H2020 project (grant number: 731011; call: H2020-EINFRA-2016-1), and the EPOS EC project (grant number: 676564; call: H2020-INFRADEV-1-2015-1)

References

[1] Bo-Christer B. & Turid H., (2004) A formalised model of the scientific publication process, Online Information Review, Vol. 28 Issue: 1, pp.8-21, doi: 10.1108/14684520410522411

[2] P. Berardino, G. Fornaro, R. Lanari, and E. Sansosti, "A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2375–2383, 2002.

[3] F. Casu, S. Elefante, P. Imperatore, I. Zinno, M. Manunta, C. De Luca, and R. Lanari, "SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 8, pp. 3285–3296, Aug. 2014.

[4] F. Casu, M. Manzo, and R. Lanari, "A quantitative assessment of the SBAS algorithm performance for surface deformation retrieval from DInSAR data," *Remote Sens. Environ.*, vol. 102, no. 3–4, pp. 195–210, Jun. 2006.

[5] I. Zinno, F. Casu, C. De Luca, S. Elefante, R. Lanari, and M. Manunta, "A Cloud Computing Solution for the Efficient Implementation of the P-SBAS DInSAR Approach," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, pp. 1–16, 2016.

[6] C. De Luca, I. Zinno, M. Manunta, R. Lanari, and F. Casu, "Large areas surface deformation analysis through a cloud computing P-SBAS approach for massive processing of DInSAR time series," *Remote Sens. Environ.*, vol. 202, pp. 3–17, 2017.

[7] C. De Luca, R. Cuccu, S. Elefante, I. Zinno, M. Manunta, V. Casola, G. Rivolta, R. Lanari, and F. Casu, "An On-Demand Web Tool for the Unsupervised Retrieval of Earth's Surface Deformation from SAR Data: The P-SBAS Service within the ESA G-POD Environment," *Remote Sens.*, vol. 7, no. 11, pp. 15630–15650, 2015.

[8] I. Zinno, L. Mossucca, S. Elefante, C. De Luca, V. Casola, O. Terzo, F. Casu, R. Lanari, "Cloud Computing for Earth Surface Deformation Analysis via Spaceborne Radar Imaging: A Case Study," in *IEEE Transactions on Cloud Computing*, vol. 4, no. 1, pp. 104-118, Jan.-March 1 2016.

[9] Manghi, P., Houssos, N., Mikulicic, M., & Jörg, B. (2012, November). The data model of the openaire scientific communication e-infrastructure. In Research Conference on Metadata and Semantic Research (pp. 168-180). Springer, Berlin, Heidelberg.

[10] Manghi, P., Bolikowski, L., Manola, N., Schirrwagen, J., & Smith, T. (2012). Openaireplus: the european scholarly communication data infrastructure. D-Lib Magazine, 18(9-10).