S-ProvFlow. Storing and Exploring Lineage Data as a Service (Postprint)

Authors: Alessandro, Spinuso, Malcolm, Atkinson, Federica, Magnoni, Alessandro, Spinuso

Date: 2022-11-28T00:00:00+00:00

Abstract

We present a set of configurable Web service and interactive tools, s-ProvFlow, for managing and exploiting records tracking data lineage during workflow runs. It facilitates detailed analysis of single executions. It helps users manage complex tasks by exposing the relationships between data, people, equipment and workflow runs intended to combine productively. Its logical model extends the PROV standard to precisely record parallel data-streaming applications. Its metadata handling encourages users to capture the application context by specifying how application attributes, often using standard vocabularies, should be added. These metadata records immediately help productivity as the interactive tools support their use in selection and bulk operations. Users rapidly appreciate the power of the encoded semantics as they reap the benefits. This improves the quality of provenance for users and management. Which in turn facilitates analysis of collections of runs, enabling users to manage results and validate procedures. It fosters reuse of data and methods and facilitates diagnostic investigations and optimisations. We present S-ProvFlow' s use by scientists, research engineers and managers as part of the DARE hyper-platform as they create, validate and use their data-driven scientific workflows.

Full Text

Preamble

S-ProvFlow: Storing and Exploring Lineage Data as a Service

Alessandro Spinuso[1]†, Malcolm Atkinson[2] & Federica Magnoni[3]
[1]Koninklijk Nederlands Meteorologisch Instituut, De Bilt, Utrecht 3731 GA, The Netherlands
[2]University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom
[3]Istituto Nazionale Geofisica e Vulcanologia, Rome, Lazio 00143, Italy

Keywords: Provenance; Productivity; Workflows; Human-in-the-loop; Visualisation

Citation: Spinuso, A., Atkinson, M., Magnoni, F.: S-ProvFlow. Storing and exploring lineage data as a service. Data Intelligence 4(2), 226-242 (2022). doi: 10.1162/dint_a_{00128}

Received: July 28, 2021; **Revised:** December 3, 2021; **Accepted:** February 4, 2022

Corresponding author: Alessandro Spinuso (Email: alessandro.spinuso@knmi.nl; ORCID: 0000-0002-0077-8491).

ABSTRACT

We present s-ProvFlow, a suite of configurable Web services and interactive tools for managing and exploiting lineage records that track data provenance during workflow executions. The system facilitates detailed analysis of individual runs and helps users manage complex tasks by exposing relationships between data, people, equipment, and workflow runs intended for productive combination. Its logical model extends the PROV standard to precisely capture parallel data-streaming applications, while its metadata handling encourages users to capture application context by specifying how attributes—often drawn from standard vocabularies—should be incorporated. These metadata records immediately enhance productivity, as the interactive tools support their use in selection and bulk operations. Users rapidly appreciate the power of encoded semantics as they reap tangible benefits, which improves provenance quality for both users and management. This, in turn, facilitates analysis of run collections, enabling users to manage results and validate procedures, fosters reuse of data and methods, and supports diagnostic investigations and optimizations. We demonstrate S-ProvFlow' s use by scientists, research engineers, and managers within the DARE hyper-platform as they create, validate, and employ their data-driven scientific workflows.

---

1. INTRODUCTION

The provenance of workflow executions, with its wealth of metadata, must be stored, processed, made comprehensible, and productively used. The S-ProvFlow system supports the acquisition and exploration of lineage and provenance data from workflows through a database, Web service, and two interactive tools. Myers et al. [1] demonstrated that productive use motivates adoption and improves metadata quality, which in turn enhances research objects as their metadata is refined through use—enabling accurate replication essential

for CWFR[1]. Users must be actively engaged through tools that improve their productivity.

While much contemporary work enables researchers to author methods abstractly, with those methods being optimally (re)mapped as technology improves while retaining user-specified semantics, we focus on the reverse information flow: from those evolving systems back to research developers and application experts. We satisfy similar demands for simplicity, comprehensibility, and stability, delivering FAIR benefits to all users by facilitating access to standardized, persistent provenance traces and the data, workflows, and components accessible via those traces. Our interactive tools demonstrate this potential, which can be exploited by many other tools, systems, and workflow technologies.

The metadata employed combines generic, widely agreed terms—such as system and software identities and geo-spatial references—with discipline- and community-specific terms representing their knowledge infrastructure, framed by general (often global) hard-won agreements ([2], page xv). As we illustrate, this is essential for grounding information in terms usable by researchers, their peers, and successors through their working practices and digital systems. There is a corresponding spectrum of persistent identification: traces judged valuable by domain experts can be allocated standard PIDs, while objects and components they touch are identified through standard or community-established and sustained methods. The refinement of metadata leading to fully described products and experiments can proceed incrementally, producing updated and refined digital objects. To achieve this, interactions between data and workflows at different maturity states should be evaluated within a collaborative and evolving ecosystem, as we demonstrate in Section 4.1. This prototypes a path that other CWFR elements must follow to gain wide adoption incrementally in research communities with substantial intellectual, technical, and political investments.

Provenance standards serve as a lingua franca encoding information gathered from multiple layers of computational workflows. Our comprehensive architecture encompasses generation, management, and access to provenance data. In previous work [3, 4], we addressed how research developers tune provenance generation by injecting metadata instructions into workflow operators via an Active provenance framework that customizes fine-grained lineage and fosters interaction between users and workflow provenance mechanisms [1]. The framework was demonstrated in the context of dispel4py [5, 6], a general-purpose analysis library for data-streaming pipelines. Depending on requirements, users can instruct dispel4py workflows to extract metadata according to a kernel of agreed terms specified by domain initiatives (as mentioned in Section 4) alongside experimental terms, which are injected into the lineage alongside general-purpose attributes characterizing the provenance model. In this follow-up paper, we present a lineage Web service and tools that develop understanding and fluency by delivering immediate benefits, showing their use for active monitoring and

retrospective analysis for diverse purposes.

S-ProvFlow exploits standard models [4, 3] to efficiently manage metadata, thereby fostering comprehension, usability, and interoperability. Parameters, intermediate outputs, and final results are automatically annotated, making traces and results discoverable and actionable while offering drill-down and FAIR access to workflow outcomes and enactment history. The system serves as the provenance service of the DARE[2] platform [7, 8], first released during the VERCE project[3]. This paper introduces its Web API and interactive tools. By managing customized provenance traces produced during workflow execution, these tools deliver understanding and control of "live" processes and encourage sharing and reuse of data and methods. We have demonstrated practical value when evaluating evidence bases and managing large numbers of runs. S-ProvFlow has provided eight years of experience working with domain experts, research developers, and support teams for sophisticated data and computation, informing our identification of vital CWFR requirements essential for quality and sustainability in multi-disciplinary professional collaborations that yield decision support we all depend on.

2. RELATED WORK

Methods for querying and visualizing provenance have been developed [9, 10, 11], addressing specific scenarios, workflow systems, or general mechanisms such as ProvStore [12]. Storage technologies are similarly specialized—for example, PBase used Neo4j[4] for the ProvONE model to enable queries on workflow traces uploaded in VisTrails XML format, supporting lineage and execution queries focusing on process involvement and data-process relationships. For our work, we chose the well-established document store MongoDB [13] to prioritize use cases accessing provenance information via data properties and process parameters. This enables S-ProvFlow' s discovery functionalities to exploit lineage produced via the Active framework [3], reflecting each user' s context and metadata. We address challenges from [14, 15]—also tackled by [10]—where provenance is annotated with rich metadata and configuration parameters relying on flexible vocabularies.

Interactive access to provenance data depends on visualization quality. Map Orbiter [16], for instance, summarizes DAGs that can be expanded on demand. S-ProvFlow instead begins with a partial visualization of the lineage graph that can be searched or browsed to show process outputs, allowing users to expand and navigate the derivation graph interactively. Alternative techniques include the Sankey diagram used by PROV-O-Viz [17] to represent flow magnitudes between activities, and radial diagrams [18] used for parallel I/O analysis [19] and pioneered by InProv [11] for provenance visualization of PASS recordings [20, 21]. Recently, this has been combined with computer graphics to improve visual efficiency [22], reducing clutter by bringing important nodes to the front.

ProvStore [12, 23] offers radial diagrams for provenance relationships. In S-ProvFlow, users compose queries to focus on specific metadata terms and values, customizing views for single computations or multiple users and runs. Platforms aiming at research artifact publication and reproducibility, such as Whole Tale [24], could benefit from adopting S-ProvFlow to complement their capability to manage preconfigured computing environments for workflow re-execution with services to monitor, review, and better represent results—for instance, by performing key provenance queries that select and render portions of overall workflow outputs.

---

3. S-PROVFLOW: ARCHITECTURE AND COMPONENTS

S-ProvFlow comprises multiple components delivering comprehensive provenance infrastructure, including a Web API and tools for interactive lineage exploration (Figure 1 [Figure 1: see original paper]). The underlying provenance model S-PROV[5] [3] builds on PROV[6] and ProvONE[7] recommendations, adding elements to encode complex lineage patterns including process delegation, distribution, and statefulness of workflow operators. In the recent DARE platform[8] deployment [25], components are organized as microservices optimized through message queue decoupling, delivering failure resilience and support for authentication infrastructures.

3.1 Lineage API

To facilitate lineage data exploitation, S-ProvFlow exposes high-level interrogation methods. We present use cases and implementation details.

Layered Workflow Activity: Workflow execution can be examined at varying detail levels, from high-level views based on classification and functional grouping of elements down to single elements with multiple processing instances. Clients can easily switch between views using MongoDB' s recommended data denormalization to exploit its powerful aggregation framework[9]. Each process invocation generates a lineage document containing detailed metadata—execution location, software characteristics, the process' s role as a higher-level function component (potentially composed of multiple operators)—alongside dynamic metadata such as execution time, data volumes, reference values, and domain metadata specified for the run or process. When queried, information is aggregated without joins, obtaining complete processing and functional information. Dynamic data is processed and aggregated to deliver the client-selected abstraction level; more aggregated documents relative to a process property yield lower information granularity [4].

Search: The API enables discovery of data and experiments through metadata term searches and semantic/functional abstractions characterizing processes. In-

tuitive metadata expressions combine value-lists and ranges to locate data and workflow executions. To accelerate searches, we use compound indexes[10] on dictionaries of the form {key:, value:}, enabling efficient queries of dynamic vocabularies without hitting index count limits.

Data Lineage: The API allows interactive navigation of data derivation graphs by specifying retrieval depth at each step. The denormalized storage approach combined with linked lists representing PROV wasDerivedFrom relationships between data entities in different documents enables easy process information retrieval during derivation navigation. Clients combine graph traversals with metadata filters to view data whose ancestors match specified properties.

Aggregations: The API provides high-level summary methods extracting comprehensive information about single or collections of runs. One method covers single-run processing dynamics, showing data transfers between processes at configurable granularity. Another reveals collaborative dynamics across runs, such as data reuse between workflows, infrastructures, and users (see Section 4.1). A final method extrapolates archive metadata, summarizing their role and occurrence for users and workflows—used in interfaces to produce personalized query term recommendations, as shown in Figure 2 [Figure 2: see original paper]. All methods leverage MongoDB' s powerful aggregation and map-reduce capabilities.

3.2 Interactive Tools

Research developers and administrators use the API for different purposes: the former validate and analyze experimental result lineage, while the latter monitor infrastructure and data exploitation by users and applications. S-ProvFlow provides tailored tools for both.

Monitoring and Validation Visualiser (MVV) assists users in fine-grain interpretation of provenance records to understand dependencies. It enables viewpoint selection and configuration through specifiable searches over domain metadata, offering data previews and dependency graph navigation. Detailed runtime diagnostics differentiate stateless and stateful processes—the latter highlighting data retention by operators such as accumulators and mergers. Figure 2 depicts the tool' s visual components showing lineage from a seismology workflow. Users search executions and data elements via simple syntax facilitating metadata searches over ranges or value lists, referencing standard vocabularies or experimental terms for specific applications. Advanced filters operate on ancestor metadata values to reduce results. Search results can be investigated interactively by browsing metadata describing products and processes. Products may be volatile, described only by metadata, but when associated with actual resources, the tool offers preview and download functions. Every search receives assistance through hints updated via incremental archive analysis using API aggregation methods (Section 3.1).

Bulk Dependency Visualiser (BDV) offers broader perspectives on com-

putational characteristics and collaborative interactions, combining radial diagrams with configurable grouping and hierarchical edge bundle techniques [22]. Users dynamically adjust viewing and grouping controls to uncover processing distribution aspects enabled by the underlying S-PROV model. The core provenance data model encompasses system-level information within application and workflow operator semantics, providing an interoperable representation of workflow resource mapping to target clusters. BDV uses these capabilities to aggregate data and accommodate views offering enactment insights at configurable detail levels, as shown in Figure 3 [Figure 3: see original paper]. Beyond single execution exploration, BDV can produce overviews of large experiments involving multiple researchers reusing and exchanging data across workflows with progressively refined metadata. This visualization could apply to evolving FDOs, foregrounding experiments whose metadata has been incrementally refined by peers to better characterize methods and results contributing to particular studies—a use case discussed in Section 4.1 for seismological applications.

---

4. TEST CASE: SEISMIC RAPID ASSESSMENT

Computational seismology faces the challenge of managing increasing volumes of recorded and simulated data downloaded from rich archives or generated by accurate, computationally sophisticated tools. Users need to easily customize available methods to adroitly explore new opportunities and meet urgent challenges. Robust provenance-driven tools are required to intelligently organize data storage and related metadata while encouraging exploration, combination, and reuse—promoting reproducibility and error detection in scientific experiments, aligning with international Earth science data handling efforts[11].

These needs become critically urgent after large seismic events, as reliable, immediate outcomes are fundamental to emergency response. Rapid assessment of seismic ground motion (RA) is a key computational seismology application embodying all aforementioned requirements and demonstrating framework applicability to different scientific contexts (e.g., [26]).

After major earthquakes, rapidly simulating seismic waveform propagation in surrounding areas and quantitatively estimating ground motion parameters is essential for impact assessment. Comparing and integrating synthetic information with recorded ground motion data improves understanding of ground response.

RA theoretical foundations and procedures are well-established, exemplified by high-level steps in Figure 4 [Figure 4: see original paper] and detailed in [3, 27, 28]. Some steps are seismology-specific while others are reusable (with adaptations) in other fields (e.g., volcanology, climate sciences; [26]). All steps require traceability and explorable metadata to support researchers in checking, reusing, and sharing methods and findings.

A fundamental RA workflow step is Waveform pre-processing, which prepares seismological data by making simulated and recorded traces consistent and comparable. This stage is used in many seismological applications (with sub-step variants) such as seismic source parameter inversion, seismic tomography, and noise cross-correlation analyses (e.g., [29, 30, 31]), with similar processing required for numerous geophysical and scientific applications (e.g., geodesy, climate sciences). Our system guarantees that results have associated provenance information about all preprocessing steps and parameterizations (Figure 2), including data properties after each step—enabling error detection and comparison of different data preparation methods and their effects on ground motion parameter estimates.

After pre-processing, RA analysis requires extraction of ground motion parameters from synthetic and recorded seismograms for comparison. Figure 5 [Figure 5: see original paper] presents the PGM (Peak Ground Motion) Parameters workflow for a single seismic station and channel. Our implementation applies combined analysis in parallel to synthetic and observed data, tracking fine-grain steps and acquired metadata. This allows researchers to trace processing in case of errors, combine and compare large data volumes, and discover intermediate results.

Drawing on strong scientific background, the Waveform pre-processing and PGM Parameters workflows benefit significantly from S-ProvFlow through proper management and description of intermediate results with usable metadata. These are typical seismological metadata adhering to recognized field standards and linkable to well-established Earth science infrastructures (e.g., [32, 33], EIDA-ORFEUS[12], FDSN[13], IRIS[14]). Additionally, user-customized metadata learned from workflow executions continuously enriches an up-to-date database. Users can check each step' s results even when outputs aren' t stored, with results and executed steps traceable in generating process and workflow context. This fosters retrieval for comparison and reuse and supports diagnosis, validation, and fine-tuning of new experimental analyses—making our metadata and provenance management approach deeply customizable for specific applications while easily adaptable to diverse scientific workflows and operational scenarios [28].

4.1 Collaborative Interactions and Metadata Refinements

The above workflow involves collaborative interactions between seismologists through data reuse across analysis stages. Figure 6 [Figure 6: see original paper] shows a radial diagram displaying runs by two users. The right half presents interlinked workflows organized into separate radiants according to conceptual tasks described by specifying concepts and metadata to contextualize involved methods. In contrast, the left side shows poor conceptual characterization typical of early exploration phases, resulting in chaotic, visually difficult-to-analyze provenance graphs. Metadata-based view customization allows users to visually restrict scope to particular data properties, suggesting reuse of some run

results or discarding those with little contribution. Combined connections and colors make these interactions evident. Interactive parameter adjustment with immediate feedback enables users to tune displays until they reveal desired information. Such visualization helps communities and research managers obtain immediate overviews of interactions across sites. Especially in the CWFR framework, it shows how study metadata has been extended and refined over time toward fully described experiments. These comprehensive diagrams significantly aid coordination of long-running or extensive research campaigns and may be used for public outreach and official reports, providing tangible perception of distributed data-intensive platform exploitation—serving another consumer category through the same provenance model.

5. CONCLUSIONS AND FUTURE WORK

Incremental improvement of provenance relevance and usefulness increases confidence in exploitation possibilities, promoting awareness of its importance. Provenance traces themselves are candidate FDOs [34], providing critical actionable information about workflows (validity, usefulness, efficiency) and each enactment' s data products—all potentially becoming FDOs depending on CWFR standards adoption, which will inevitably be an incremental process following paths we have pioneered. Through expert participation, provenance is ready-made for validation and results management use cases.

We explored technical solutions and standard models for next-generation WMSs [35] encompassing FDO challenges concerning evaluation and traceability of experimental results. This depended on co-design and co-development with domain experts in communities with long-established global knowledge infrastructure, corresponding standards, and agreed practices. This paper focused on computational seismologists; a companion paper [36] reports how provenance management empowers reproducibility of interactive workspaces beyond workflows, including climate science use cases. Edwards [2] clarifies the extensive, complex knowledge infrastructure that took over a century to develop incrementally. Our standards and technologies must prove value alongside established systems before penetrating working practices.

Improving provenance quality and use by delivering immediate benefits to broad user ranges will encourage adoption and engagement. Enhanced productivity and decision support quality will motivate investment in sustainably collecting and preserving vital provenance records with sufficient content, yielding long-term benefits for research procedure quality and evidence underpinning life-critical decisions. We adopted services and tools demonstrating effectiveness, providing easy access, visualization, and navigation through provenance information and metadata with direct links to physical data resources. Researchers exploit these tools to improve science by quickly detecting and solving anomalies and optimizing multi-run and data combinations for complex appli-

cations. Research engineers and developers are facilitated in improving resource and data exploitation (Section 4.1). We illustrated these capabilities through real seismology application use cases.

Future work will address and improve preliminary results [25] enabling CWL-Prov import into s-ProvFlow to scale benefits to broader WMS collections. Similar interdisciplinary co-development demonstrating immediate benefits will test and develop other canonical workflow technology aspects, incentivizing widespread experimental adoption and sustained quality, capability, and adoption growth. We anticipate collaborations embedding in many research contexts to improve standards including provenance, develop enabled tools and work environments, and build adoption momentum—delivering more FDOs corresponding to all supported research aspects and extending FDO standards deeper into established practices.

AUTHOR CONTRIBUTIONS

A. Spinuso (spinuso@knmi.nl) and M. Atkinson (Malcolm.Atkinson@ed.ac.uk) contributed to system research, design, and implementation. F. Magnoni (federica.magnoni@ingv.it) verified capabilities by implementing a use case leveraging the framework. All authors contributed to manuscript writing.

ACKNOWLEDGEMENTS

This work was supported by the EU FP7-Infrastructure project VERCE (No. 283543) and EU H2020 project DARE (No. 777413). These projects comprise large teams of software engineers and researchers in Climate and Earth Sciences who contributed to implementing and adopting the illustrated technologies. We thank them for continuous support and proactive participation.

REFERENCES

[1] Myers, J., et al.: Towards sustainable curation and preservation: The sead project' s data services approach. In: 2015 IEEE 11th International Conference on e-Science, pp. 485–494 (2015)

[2] Edwards, P.N.: A vast machine: Computer models, climate data, and the politics of global warming. MIT Press, Cambridge (2010)

[3] Spinuso, A., Atkinson, M., Magnoni, F.: Active provenance for data-intensive workflows: Engaging users and developers. In: 2019 15th International Conference on eScience (eScience), pp. 560–569 (2019)

[4] Spinuso, A.: Active provenance for data intensive research. PhD dissertation, The University of Edinburgh (2018). Available at: http://hdl.handle.net/1842/33181. Accessed 4 February 2022

[5] Filgueira, R., et al.: dispel4py: An agile framework for data-intensive escience. In: The 11th IEEE International Conference on e-Science, pp. 454-464 (2015)

[6] Filguiera, R., et al.: dispel4py: A python framework for data-intensive scientific computing. In 2014 International Workshop on Data Intensive Scalable Computing Systems, pp. 9-16 (2014)

[7] Klampanos, I., et al.: Dare: A reflective platform designed to enable agile data-driven research on the cloud. In: The 15th International Conference on eScience (eScience), No. 19473092 (2019)

[8] Atkinson, M., et al: Comprehensible control for researchers and developers facing data challenges. In: The 15th International Conference on eScience (eScience), No. 19473140 (2019)

[9] Hunter, J., Cheung, K.: Provenance explorer—A graphical interface for constructing scientific publication packages from provenance trails. International Journal on Digital Libraries 7, 99-107(2007)

[10] Cuevas-Vicenttín, V., et al.: The PBase scientific workflow provenance repository. International Journal of Digital Curation 9(2), 28-38 (2014)

[11] Borkin, M.A., et al.: Evaluation of filesystem provenance visualization tools. IEEE Transactions on Visualization and Computer Graphics 19(12), 2476-2485 (2013)

[12] Huynh, T.D., Moreau, L.: Provstore: A public provenance repository. In: International Provenance and Annotation Workshop, pp. 275-277 (2014)

[13] MongoDB Document-oriented Data Base (2020). Available at: http://mongodb.org. Accessed 1 February 2022

[14] De Oliveira, D., Silva, V., Mattoso, M.: How much domain data should be in provenance databases? In: Proceedings of the 7th USENIX Conference on Theory and Practice of Provenance (TaPP' 15), pp. 1-9 (2015)

[15] Gadelha, L.M.R., et al.: MTCProv: A practical provenance query framework for many-task scientific computing. Distributed Parallel Databases 30, 351-370 (2012)

[16] Seltzer, M., Macko, P.: Provenance map orbiter: Interactive exploration of large provenance graphs. In: Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP' 11), pp. 20-21(2011)

[17] Hoekstra, R, Groth, P: Prov-O-Viz—understanding the role of activities in provenance. In: Ludäscher, B., Plale, B. (eds.) Provenance and Annotation

of Data and Processes, pp. 215–220. Springer International Publishing, Cham (2015)

[18] Draper, G.M., Livnat, Y., Riesenfeld, R.F.: A survey of radial methods for information visualization. IEEE Transactions on Visualization and Computer Graphics 15(5), 759–776 (2009)

[19] Sigovan, C., et al.: A visual network analysis method for large-scale parallel I/O systems. In: IEEE 27th International Symposium on Parallel Distributed Processing (IPDPS), pp. 308–319 (2013)

[20] PASS: Provenance-aware storage systems. Available at: http://www.eecs.harvard.edu/syrah/pass/. Accessed 1 February 2022

[21] Muniswamy-Reddy, K.-K., et al.: Provenance-aware storage systems. In: USENIX Annual Technical Conference, General Track, pp. 43–56 (2006)

[22] Holten, D.: Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Transactions on Visualization and Computer Graphics 12(5), 741–748 (2006)

[23] ProvStore, provenance storage and distribution. Available at: https://openprovenance.org/store/. Accessed 1 February 2022

[24] Brinckman, A., et al.: Computing environments for reproducibility: Capturing the "whole tale" . Future Generation Computer Systems 94, 854–867 (2019)

[25] Spinuso, A.: D3.4 data lineage services ii (2020). Available at: https://doi.org/10.5281/zenodo.4738814. Accessed 1 February 2022

[26] Malcolm, A., et al.: Dare architecture and technology D2.2 (Version 1) (2020). Available at: https://doi.org/10.5281/zenodo.4733801. Accessed 1 February 2022

[27] Magnoni, F., et al.: D6.3 pilot tools and services, execution and evaluation report I (2019). Available at: https://doi.org/10.5281/zenodo.4739217. Accessed 1 February 2022

[28] Magnoni, F., et al.: D6.4 pilot tools and services, execution and evaluation report II (2020). Available at: https://doi.org/10.5281/zenodo.4740027. Accessed 1 February 2022

[29] Scognamiglio, L., et al.: Uncertainty estimations for moment tensor inversions: The issue of the 2012 May 20 Emilia earthquake. Geophysical Journal International 206(2), 792–806 (2016)

[30] Bozdag˘, E,. et al.: Global adjoint tomography: First-generation model. Geophysical Journal International 207(3), 1739–1766 (2016)

[31] Zaccarelli, L., et al.: Variations of crustal elastic properties during the 2009 L' Aquila earthquake inferred from cross-correlations of ambient seismic noise. Geophysical Research Letters 38(24), L24304 (2011)

[32] Danecek, P., et al.: The Italian node of the European integrated data archive. Seismological Research Letters 92(3), 1726–1737 (2021)

[33] Michelini, A., et al.: INSTANCE—The Italian seismic dataset for machine learning. Earth System Science Data Discuss Preprint (2021). Available at: https://doi.org/10.5194/essd-2021-164. Accessed 1 February 2022

[34] De Smedt, K., Koureas, D., Wittenburg, P.: FAIR digital objects for science: From data pieces to actionable knowledge units. Publication 8(2), Article No. 21 (2020)

[35] Deelman, E., et al.: The future of scientific workflows. The International Journal of High Performance Computing Applications 32(1), 159–175 (2018)

[36] Spinuso, A., et al.: SWIRRL: Managing provenance-aware and reproducible workspaces. Data Intelligence 4(2), 243–258 (2022)

AUTHOR BIOGRAPHY

Alessandro Spinuso is a researcher at the RD Observations and Data Technology division of the Royal Netherlands Meteorological Institute (KNMI). He earned his Ph.D. in Computer Science at the University of Edinburgh (UK) in 2017. At KNMI, he serves as Researcher and Product Owner within an Agile RD team developing Provenance-aware Data Analysis services. His main research interest is management and exploitation of provenance information in user-controlled computational environments providing notebooks and workflow systems for data-intensive analysis. He participates in several EU initiatives (H2020, Copernicus), focusing on e-Science infrastructure development for Earth Science research in Europe (EPOS, ENVRIFair, IS-ENES3, DARE, and C3S). More recently, he is an invited expert to the IPCC TG-Data, a working group dedicated to FAIR management of data and methods for upcoming IPCC reports. ORCID: 0000-0002-0077-8491

Malcolm Atkinson, Ph.D., FBCS, FRSE, has spent 53 years improving capacity through education and research to enable individuals, organizations, and society to make best use of their data. He leads the data-intensive research group focusing on socio-technical architectures to address this challenge. He is Professor of e-Science in the Artificial Intelligence Applications Institute, School of Informatics, University of Edinburgh. ORCID: 0000-0003-2632-0013

Federica Magnoni is a researcher in computational seismology at Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome, Italy. She earned her Ph.D. in Geophysics at the University of Bologna (Italy) in 2012. She has substantial experience in 3D seismic wave propagation forward and inverse problems and ground shaking assessment, exploiting numerical methods for realistic 3D earth models and point or finite earthquake source models. She participated in numerous projects exploiting national and international HPC resources for HPC-

and data-intensive applications (PRACE and ISCRA projects). She worked and is presently involved in several EU projects (FP7, H2020), focusing on e-Science infrastructure development for Earth Science (VERCE, EPOS-IP, and DARE), scientific applications with flagship European codes prepared for pre-Exascale and Exascale computations (ChEESE), and national projects on earthquake near real-time simulation and social media role in emergency communication (PRIN 2012–SHAKEnetworks), or machine learning exploitation for seismological applications (Pianeta Dinamico - SOME). ORCID: 0000-0002-0833-8044

Note: Figure translations are in progress. See original paper for figures.