2021

Istituto di Scienza e Tecnologie
dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Annual Reports

InfraScience Research Activity Report 2020

InfraScience lab., CNR-ISTI, Pisa, Italy

ISTI-AR-2021/002

Istituto di Scienza e Tecnologie
dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

InfraScience Research Activity Report 2020
InfraScience lab.
ISTI-AR-2021/002

Abstract
InfraScience is a research group of the National Research Council of Italy - Institute of Information Science and Technologies (CNR - ISTI) based in Pisa, Italy. This report documents the research activity performed by this group in 2020 to highlight the major results. In particular, the InfraScience group confronted with research challenges characterising Data Infrastructures, e\-Sci\-ence, and Intelligent Systems. The group activity is pursued by closely connecting research and development and by promoting and supporting open science. In fact, the group is leading the development of two large scale infrastructures for Open Science, \ie D4Science and OpenAIRE. During 2020 InfraScience members contributed to the publishing of 30 papers, to the research and development activities of 12 research projects (11 funded by EU), to the organization of conferences and training events, to several working groups and task forces.

Infrastructure, Open Science, Intelligent Systems, Activity Report, Research Report

Citation
InfraScience lab. *InfraScience Research Activity Report 2020*, ISTI Annual Reports 2021/002. DOI: 10.32079/ISTI-AR-2021/002

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

ISTI-AR-2021/002

Technical Report, December 2021
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

InfraScience Research Activity Report 2020

Michele Artini, Massimiliano Assante, Claudio Atzori, Miriam Baglioni, Alessia Bardi, Leonardo Candela, Giovanni Casini, Donatella Castelli*, Roberto Cirillo, Gianpaolo Coro, Franca Debole, Andrea Dell'Amico, Luca Frosini, Sandro La Bruzzo, Emma Lazzeri, Lucio Lelii, Paolo Manghi, Francesco Mangiacrapa, Andrea Mannocci, Pasquale Pagano, Giancarlo Panichi, Tommaso Piccioli, Fabio Sinibaldi, Umberto Straccia

Abstract
InfraScience is a research group of the National Research Council of Italy - Institute of Information Science and Technologies (CNR - ISTI) based in Pisa, Italy. This report documents the research activity performed by this group in 2020 to highlight the major results. In particular, the InfraScience group confronted with research challenges characterising Data Infrastructures, eScience, and Intelligent Systems. The group activity is pursued by closely connecting research and development and by promoting and supporting open science. In fact, the group is leading the development of two large scale infrastructures for Open Science, i.e., D4Science and OpenAIRE. During 2020 InfraScience members contributed to the publishing of 30 papers, to the research and development activities of 12 research projects (11 funded by EU), to the organization of conferences and training events, to several working groups and task forces.

Keywords
Infrastructure — Open Science — Intelligent Systems

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy
***Corresponding author**: donatella.castelli@isti.cnr.it

Contents

1. Introduction

Modern science is heavily data and compute-intensive, AI-assisted, participatory, and multidisciplinary. Sharing and publishing scientific results are activities going to be revolutionised to support openness, transparency and reproducibility, and to enable rewards for scientists who publish results of their work beyond the scientific articles. These approaches are expressions of a profound evolution of science practices that on the one hand is enacted by, and on the other demand for, a continuous innovation in IT instruments and approaches.

InfraScience is a research group working to contribute to this evolution by investigating, experimenting, and closely connecting research and development of innovative digital infrastructures, information systems, and smart solutions for fostering and empowering data-centered research. InfraScience is a research group of the National Research Council of Italy - Institute of Information Science and Technologies (CNR - ISTI)[1] based in Pisa, Italy. It consists of 27 members: 21 research staff and 6 technical staff. Moreover, it counts on 16 collaborators including postdocs, doctoral students, and research associates.

This report documents the research activity performed by the group in 2020, the resulting publications, the active research projects, and the services and infrastructures operated. In particular, Sec. 2 describes the topics characterising InfraScience research. Sec. 3 reports on the publications pro-

[1] www.isti.cnr.it

duced by the group. Sec. 4 documents the research projects InfraScience contributed to. Sec. 5 describes the major developments of the two infrastructures the team is responsible for. Sec. 6 reports on the software artefacts released by InfraScience. Sec. 7 reports on the organised events. Sec. 8 details the training activity performed by InfraScience. Sec. 9 documents the working groups and task forces InfraScience members participate in. Finally, Sec. 10 concludes the report and give prospects on future research activities.

2. Research topics

The research activities conducted by infraScience members revolves around three major topics: Data Infrastructures, e-Science, and Intelligent Systems.

2.1 Data infrastructures

This is a very broad research area including models, approaches and solutions underlying the development and operation of data infrastructures suitable for thematic and interdisciplinary scientific contexts characterized by variability, heterogeneity, reusability and presence of "big data". The group is confronting with these challenges by closely connecting research and development. In fact, InfraScience is responsible for the development of two large scale infrastructures supporting open science, namely D4Science and OpenAIRE cf. Sec. 5. The major themes and investigations include approaches and solutions for the delivery of Virtual Research Environments and Science Gateways for various communities of practice, e.g., [2, 22], design and development activities of infrastructure-oriented solutions for data analytics at scale, e.g., [9], natural language processing, e.g., [18], scholarly communication graphs management, e.g., [27].

2.2 eScience

This is a wide research domain including models, approaches and solutions to carry out collaborative data-driven and reproducible analytical workflows while supporting, at the same time, sharing, publishing, validation, and monitoring (usage and impact) of the related scientific outcomes (publications, datasets, software, etc.). The group is confronting with several challenges belonging to the domain including the study of problems and solution regarding the implementation and diffusion of open science practices e.g., [28, 31], the implementation of open science practices supported by data infrastructures in specific investigations, e.g., [15, 19, 20].

2.3 Intelligent Systems

This research area concerns AI-assisted methods and approaches to enable humans and systems to discover, access, process, and learn structured and unstructured information. InfraScience is confronting with challenges including the development of tools and solutions for managing ontologies, e.g., [21], knowledge representation challenges, e.g., [32], defeasible-reasoning, e.g., [7], design and development of discovery services, e.g., [5], the development of domain specific information systems, e.g., [30].

3. Papers

The following papers have been published by InfraScience members in collaboration with researchers from several Institutions and scientific disciplines. In particular, InfraScience contributed 11 articles in journals, 9 papers to conferences, 4 chapters in books, and 7 publications including technical reports and other papers.

3.1 Contribution to Journals

InfraScience members contributed to the following papers published in journals.

Fudge: Fuzzy ontology building with consensuated fuzzy datatypes [21] authored by U. Straccia for the Fuzzy Sets and Systems.

Summary: An important problem in Fuzzy OWL 2 ontology building is the definition of fuzzy membership functions for real-valued fuzzy sets (so-called fuzzy datatypes in Fuzzy OWL 2 terminology). In this paper, we present a tool, called Fudge, whose aim is to support the consensual creation of fuzzy datatypes by aggregating the specifications given by a group of experts. Fudge is freeware and currently supports several linguistic aggregation strategies, including the convex combination, linguistic OWA, weighted mean and fuzzy OWA, and easily allows to build others in. We also propose and have implemented two novel linguistic aggregation operators, based on a left recursive form of the convex combination and of the linguistic OWA.

Predicting geographical suitability of geothermal power plants [19] authored by G. Coro for the Journal of Cleaner Production.

Summary: A large and increasing number of countries use geothermal energy as power source for domestic and industrial applications. Geothermal power plants produce energy out of this natural and renewable source in a sustainable way and contribute to reduce global warming. However, power plants effectiveness depends on the suitability of an area to geothermal energy production, which is a complex and unknown combination of many environmental factors. Nowadays, geothermal suitability assessments require invasive inspections, high costs, and legal permissions. Thus, having a global suitability map of geothermal sites as reference would be useful prior knowledge during assessments, and would help saving time and money. In this paper, the first suitability map of potential geothermal sites at global scale is presented. The map is the result of the application of data collection and preparation processes, and a Maximum Entropy model, to geospatial data potentially correlated with geothermal site suitability and geothermal plants operation. The reliability of our map is assessed against currently active and planned geothermal power plants. Our approach follows the Open Science paradigm that guarantees results reproduction and transparency, and allows stakeholders to reuse the produced standardised data, services, and Web interfaces in other experiments or to generate new maps at regional scale. Overall, our results can help scientists, industry operators, and policy makers in geothermal sites assessments. Also, our approach supports communication with citizens whose territories are involved in probing and assessments, in order to transparently inform them about the reasons driving the selection of their territory and the potential future benefits.

A global-scale ecological niche model to predict SARS-CoV-2 coronavirus infection rate [15] authored by G. Coro for Ecological Modelling.

Summary: COVID-19 pandemic is a global threat to human health and economy that requires urgent prevention and monitoring strategies. Several models are under study to control the disease spread and infection rate and to detect possible factors that might favour them, with a focus on understanding the correlation between the disease and specific geophysical parameters. However, the pandemic does not present evident environmental hindrances in the infected countries. Nevertheless, a lower rate of infections has been observed in some countries, which might be related to particular population and climatic conditions. In this paper, infection rate of COVID-19 is modelled globally at a 0.5°resolution, using a Maximum Entropy-based Ecological Niche Model that identifies geographical areas potentially subject to a high infection rate. The model identifies locations that could favour infection rate due to their particular geophysical (surface air temperature, precipitation, and elevation) and human-related characteristics (CO2 and population density). It was trained by facilitating data from Italian provinces that have reported a high infection rate and subsequently tested using datasets from World countries' reports. Based on this model, a risk index was calculated to identify the potential World countries and regions that have a high risk of disease increment. The distribution outputs foresee a high infection rate in many locations where real-world disease outbreaks have occurred, e.g. the Hubei province in China, and reports a high risk of disease increment in most World countries which have reported significant outbreaks (e.g. Western U.S.A.). Overall, the results suggest that a complex combination of the selected parameters might be of integral importance to understand the propagation of COVID-19 among human populations, particularly in Europe. The model and the data were distributed through Open-science Web services to maximise opportunities for re-usability regarding new data and new diseases, and also to enhance the transparency of the approach and results.

Entity deduplication in big data graphs for scholarly communication [27] authored by P. Manghi, C. Atzori, M. De Bonis, and A. Bardi for the Data Technologies and Applications journal.

Summary: Several online services offer functionalities to access information from "big research graphs" (e.g. Google Scholar, OpenAIRE, Microsoft Academic Graph), which correlate scholarly/scientific communication entities such as publications, authors, datasets, organizations, projects, funders, etc. Depending on the target users, access can vary from search and browse content to the consumption of statistics for monitoring and provision of feedback. Such graphs are populated over time as aggregations of multiple sources and therefore suffer from major entity-duplication problems. Although deduplication of graphs is a known and actual problem, existing solutions are dedicated to specific scenarios, operate on flat collections, local topology-drive challenges and cannot therefore be re-used in other contexts. This work presents GDup, an integrated, scalable, general-purpose system that can be customized to address deduplication over arbitrary large information graphs. The paper presents its high-level architecture, its implementation as a service used within the OpenAIRE infrastructure system and reports numbers of real-case experiments. GDup provides the functionalities required to deliver a fully-fledged entity deduplication workflow over a generic input graph. The system offers out-of-the-box Ground Truth management, acquisition of feedback from data curators and algorithms for identifying and merging duplicates, to obtain an output disambiguated graph. To our knowledge GDup is the only system in the literature that offers an integrated and general-purpose solution for the deduplication graphs, while targeting big data scalability issues. GDup is today one of the key modules of the OpenAIRE infrastructure production system, which monitors Open Science trends on behalf of the European Commission, National funders and institutions.

Global Food-source Identifier (GFI): Collaborative virtual research environment and shared data catalogue for the foodborne outbreak investigation international community [30] authored by L. Candela for the Food Control journal.

Summary: The source of a foodborne disease outbreak (FBO) is often difficult to identify, especially in the early phase where interventions would be most efficient. In addition, data on FBOs are mostly scattered in different formats either in national databases and reports or within pathogen-specific or regional reporting networks, both of which are often only accessible to a selected number of individuals. Here, we developed an international, open, shared and searchable data catalogue of past FBOs – the Global Food-source Identifier (GFI). GFI was developed with two objectives: a) to create a collaborative online community of FBO investigators, encouraging the international sharing of data in a harmonized, detailed and comparable manner and b) to support foodborne outbreak investigation worldwide by providing access to detailed records of past outbreaks, which can convey valuable insight into potential 'risk foods' of a detected pathogen. GFI is hosted within a Virtual Research Environment (VRE), which offers additional features to facilitate the collaboration between the outbreak investigators. These features allow document exchange, communication and data visualization and analysis between the VRE members. Based on scientific literature on foodborne outbreaks and discussions within a working group, we selected a total of 46 attributes characterising the outbreak records to be included in the catalogue, aggregated under the four overarching categories causative agent, epidata, food source and report details. Detailed descriptions of the attributes in the catalogue and instructions for harmonized data reporting are available on a wiki page in the VRE. At the time of writing and public launch of GFI, the data catalogue was populated with records of 102 FBOs occurred in Denmark over a period of 12 years (2005–2016) and covering the most frequent pathogens and a broad range of typing methods. The VRE features that enable data analysis, document sharing and communication between members were applied for the graphical representation of the records available in GFI, and for the sharing of results and script files within the VRE. The descriptive analysis included the relationship between the most frequent causative agents and outbreak food sources. Such results can support a risk-based food sampling strategy in the very beginning of a foodborne outbreak investigation. The Global Food-source Identifier is a data catalogue specifically designed to host an international collection of FBO records reported in a detailed and harmonized manner. It is implemented in a virtual research environment that offers key features to facilitate and enhance the global collaboration and data sharing among FBO investigators. Once in active use by the international food safety community, we envisage that GFI will contribute to the success of FBO investigations worldwide.

NLPHub: An e-Infrastructure-based text mining hub [18] authored by G. Coro, G. Panichi, and P. Pagano for the Concurrency and Computation: Practice and Experience journal.

Summary: Text mining involves a set of processes that analyze text to extract high-quality information. Among its large number of applications, there are experiments that tackle big data challenges using complex system architectures. However, text mining approaches are neither easy to discover and use nor easily combinable by end-users. Furthermore, they should be contextualized within new approaches to science (eg, Open Science) that ensure longevity and reuse of methods and results. This article presents NLPHub, a distributed system that orchestrates and combines several state-of-the-art text mining services that recognize spatiotemporal events, keywords, and a large set of named entities. NLPHub adopts an Open Science approach, which fosters the reproducibility, repeatability, and reusability of methods and results, by using an e-Infrastructure supporting data-intensive Science. NLPHub adds Open Science-compliance to the connected services through the use of representational standards for services and computations. It also manages heterogeneous service access policies and enables collaboration and sharing facilities. This article reports a performance assessment based on an annotated corpus of named entities, which demonstrates that NLPHub can improve the performance of the single-integrated processes by cleverly combining their output.

Open science and artificial intelligence supporting blue growth [16] authored by G. Coro for the Environmental Engineering and Management Journal.

Summary: The long-term EU strategy to support the sustainable growth of the marine and maritime sectors (Blue Growth) involves economic and ecological topics that call for new computer science systems to produce new knowledge after processing large amounts of data (Big Data), collected both at academic and industrial levels. Today, Artificial Intelligence (AI) can satisfy the Blue Growth strategy requirements by managing Big Data, but requires effective multi-disciplinary interaction between scientists. In this context, new Science paradigms, like Open Science, are born to promote the creation of computational systems to process Big Data while supporting collaborative experimentation, multi-disciplinarity, and the re-use, repetition, and reproduction of experiments and results. AI can use Open Science systems by making domain and data experts cooperate both between them and with AI modellers. In this paper, we present examples of combined AI and Open Science-oriented applications in marine science. We explain the direct benefits these bring to the Blue Growth strategy and the indirect advantages deriving from their re-use in other applications than their originally intended ones.

Realizing virtual research environments for the agri-food community: The AGINFRA PLUS experience [2] authored by M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi and F. Sinibaldi for the Concurrency and Computation: Practice and Experience journal.

Summary: The enhancements in IT solutions and the open science movement are injecting changes in the practices dealing with data collection, collation, processing, analytics, and publishing in all the domains, including agri-food. However, in implementing these changes one of the major issues faced by the agri-food researchers is the fragmentation of the "assets" to be exploited when performing research tasks, for example, data of interest are heterogeneous and scattered across several repositories, the tools modelers rely on are diverse and often make use of limited computing capacity, the publishing practices are various and rarely aim at making available the "whole story" including datasets, processes, and results. This paper presents the AGINFRA PLUS endeavor to overcome these limitations by providing researchers in three designated communities with Virtual Research Environments facilitating the use of the "assets" of interest and promote collaboration.

Principles of KLM-style Defeasible Description Logics [7] authored by G. Casini for the ACM Transactions on Computational Logic.

Summary: The past 25 years have seen many attempts to introduce defeasible-reasoning capabilities into a description logic setting. Many, if not most, of these attempts are based on preferential extensions of description logics, with a significant number of these, in turn, following the so-called

KLM approach to defeasible reasoning initially advocated for propositional logic by Kraus, Lehmann, and Magidor. Each of these attempts has its own aim of investigating particular constructions and variants of the (KLM-style) preferential approach. Here our aim is to provide a comprehensive study of the formal foundations of preferential defeasible reasoning for description logics in the KLM tradition. We start by investigating a notion of defeasible subsumption in the spirit of defeasible conditionals as studied by Kraus, Lehmann, and Magidor in the propositional case. In particular, we consider a natural and intuitive semantics for defeasible subsumption, and we investigate KLM-style syntactic properties for both preferential and rational subsumption. Our contribution includes two representation results linking our semantic constructions to the set of preferential and rational properties considered. Besides showing that our semantics is appropriate, these results pave the way for more effective decision procedures for defeasible reasoning in description logics. Indeed, we also analyse the problem of non-monotonic reasoning in description logics at the level of entailment and present an algorithm for the computation of rational closure of a defeasible knowledge base. Importantly, our algorithm relies completely on classical entailment and shows that the computational complexity of reasoning over defeasible knowledge bases is no worse than that of reasoning in the underlying classical DL ALC.

Data sources and persistent identifiers in the Open Science Research Graph of OpenAIRE [31] authored by A. Bardi and P. Manghi for the International journal of digital curation.

Summary: In this article, we give an overview of the data source typologies used in OpenAIRE and provide an outline on the role of persistent identifiers in the aggregation, curation and provision workflows that lead to the generation of the Research Graph in OpenAIRE.

Measuring success for a future vision: Defining impact in science gateways/virtual research environments [8] authored by L. Candela for the Concurrency and Computation: Practice and Experience journal.

Summary: Scholars worldwide leverage science gateways/-virtual research environments (VREs) for a wide variety of research and education endeavors spanning diverse scientific fields. Evaluating the value of a given science gateway/VRE to its constituent community is critical in obtaining the financial and human resources necessary to sustain operations and increase adoption in the user community. In this article, we feature a variety of exemplar science gateways/VREs and detail how they define impact in terms of, for example, their purpose, operation principles, and size of user base. Further, the exemplars recognize that their science gateways/VREs will continuously evolve with technological advancements and standards in cloud computing platforms, web service architectures, data management tools and cybersecurity. Correspondingly, we present a number of technology advances that could be incorporated in next-generation science gateways/VREs to enhance their scope and scale of their operations for greater success/impact. The exemplars are selected from owners of science gateways in the Science Gateways Community Institute (SGCI) clientele in the United States, and from the owners of VREs in the International Virtual Research Environment Interest Group (VRE-IG) of the Research Data Alliance. Thus, community-driven best practices and technology advances are compiled from diverse expert groups with an international perspective to envisage futuristic science gateway/VRE innovations.

3.2 Contribution to Conferences

InfraScience members contributed to the following papers presented in international and national conferences.

RepOSGate: Open Science Gateways for Institutional Repositories [1] authored by M. Artini, L. Candela, and P. Manghi for the 16th Italian Research Conference on Digital Libraries, IRCDL 2020, Bari, Italy, January 30-31, 2020.

Summary: Most repository platforms used to operate Institutional Repositories fail at delivering a complete set of functionalities required by institutions and researchers to fully comply with Open Science publishing practices. This paper presents RepOSGate, a software that implements an overlay application capable of collecting metadata records from a repository and transparently deliver search, statistics, upload of Open Access versions functionalities over an enhanced version of the metadata collection, which include: links to datasets, Open Access versions of the artifacts, links to projects from several funders, subjects, citations, etc. The paper will also present two instantiations of RepOSGate, used to enhance the publication metadata collections of two CNR institutes: Institute of Information Science and Technologies (ISTI) and Institute of Marine Sciences (ISMAR).

Using virtual research environments in agro-environmental research [25] authored by L. Candela for the 13th International Symposium on Environmental Software Systems, ISESS 2020, Wageningen, The Netherlands, 5-7 February 2020.

Summary: Tackling some of the grand global challenges, agro-environmental research has turned more and more into an international venture, where distributed research teams work together to solve complex research questions. Moreover, the interdisciplinary character of these challenges requires that a large diversity of different data sources and information is combined in new, innovative ways. There is a pressing need to support researchers with environments that allow them to efficiently work together and co-develop research. As research is often data-intensive, and big data becomes a common part of a lot of research, such environments should also offer the resources, tools and workflows that allow to process data at scale if needed. Virtual research environments (VRE), which combine working in the Cloud, with collaborative functions and state of the art data science tools, can be a potential solution. In the H2020 AGINFRA+ project, the usability of the VREs has been explored for use cases around agro-climatic modelling. The implemented pilot application for crop growth modelling has successfully shown that VREs can support distributed research teams in co-development, helps them to adopt open science and that the VRE's cloud computing facilities allow large scale modelling applications.

AGINFRA PLUS: running crop simulations on the D4Science distributed e-Infrastructure [23] authored by L. Candela for the 13th International Symposium on Environmental Software Systems, ISESS 2020, Wageningen, The Netherlands, 5-7 February 2020.

Summary: Virtual Research Environments (VREs) bridge the gap between the compute and storage infrastructure becoming available as the 'cloud', and the needs of researchers for tools supporting open science and analytics on ever larger datasets. In the AGINFRA PLUS project such a VRE, based on the D4Science platform, was examined to improve and test its capabilities for running large numbers of crop simulations at field level, based on the WOFOST-WISS model and Dutch input datasets from the AgroDataCube. Using the gCube DataMiner component of the VRE, and based on the Web Processing Service standard, a system has been implemented that can run such workloads successfully on an available cluster, and with good performance, providing summarized results to agronomists for further analysis. The methods used and the resulting implementation are briefly described in this paper. Overall the approach seems viable and opening the door to many follow-up implementation opportunities and further research. Some of them are indicated in more detail in the conclusions.

Context-Driven Discoverability of Research Data [5] authored by M. Baglioni, P. Manghi, and A. Mannocci for the 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020.

Summary: Research data sharing has been proved to be key for accelerating scientific progress and fostering interdisciplinary research; hence, the ability to search, discover and reuse data items is nowadays vital in doing science. However, research data discovery is yet an open challenge. In many cases, descriptive metadata exhibit poor quality, and the ability to automatically enrich metadata with semantic information is limited by the data files format, which is typically not textual and hard to mine. More generally, however, researchers would like to find data used across different research experiments or even disciplines. Such needs are not met by traditional metadata description schemata, which are designed to freeze research data features at deposition time. In this paper, we propose a methodology that enables "context-driven discovery" for research data thanks to their proven usage across research activities that might differ from the original one, potentially across diverse disciplines. The methodology exploits the collection of publication-dataset and dataset-dataset links provided by OpenAIRE Scholexplorer data citation index so to propagate articles metadata into related research datasets by leveraging semantic relatedness. Such "context propagation" process enables the construction of "context-enriched" metadata of datasets, which enables "context-driven" discoverability of research data. To this end, we provide a real-case evaluation of this technique applied to Scholexplorer. Due to the broad coverage of Scholexplorer, the evaluation documents the effectiveness of this technique at improving data discovery on a variety of research data repositories and databases.

How Much Knowledge is in a Knowledge Base? Introducing Knowledge Measures (Preliminary Report) [32] authored by U. Straccia for the 24th European Conference on Artificial Intelligence, ECAI 2020, Santiago de Compostela, Spain, 29 August - 8 September 2020.

Summary: In this work we address the following question: can we measure how much knowledge a knowledge base represents? We answer to this question (*i*) by describing properties (axioms) that a knowledge measure we believe should have in measuring the amount of knowledge of a knowledge base (kb); and (*ii*) provide a concrete example of such a measure, based on the notion of entropy. We also introduce related kb notions such as (*i*) accuracy; (*ii*) conciseness; and (*iii*) Pareto optimality. Informally, they address the following questions: (*i*) how precise is a kb in describing the actual world? (*ii*) how succinct is a kb w.r.t. the knowledge it represents? and (*iii*) can we increase accuracy without decreasing conciseness, or vice-versa?

Open Science Observatory: Monitoring Open Science in Europe [28] authored by P. Manghi for the International Workshop on Assessing Impact and Merit in Science (AIMinScience 2020), in conjunction with the 24th International Conference on Theory and Practice of Digital Libraries (TPDL 2020).

Summary: Monitoring and evaluating Open Science (OS) practices and research output in a principled and continuous way is recognised as one of the necessary steps towards its wider adoption. This paper presents the Open Science Observatory, a prototype online platform which combines data gathered from OpenAIRE e-Infrastructure and other public data sources and informs users via rich visualizations on different OS indicators in Europe.

Rational Defeasible Belief Change [12] authored by G. Casini for the 17th International Conference on Principles of Knowledge Representation and Reasoning.

Summary: We present a formal framework for modelling belief change within a nonmonotonic reasoning system. Belief change and non-monotonic reasoning are two areas that are formally closely related, with recent attention being paid towards the analysis of belief change within a non-monotonic environment. In this paper we consider the classical AGM belief change operators, contraction and revision, applied to a defeasible setting in the style of Kraus, Lehmann, and Magidor. The investigation leads us to the consideration of the problem of iterated change, generalising the classical work of Darwiche and Pearl. We characterise a family of operators for iterated revision, followed by an analogous characterisation of operators for iterated contraction. We start considering belief change operators aimed at preserving logical consistency, and then characterise analogous operators aimed at the preservation of coherence – an important notion within the field of logic-based ontologies.

BKLM - An expressive logic for defeasible reasoning [29] authored by G. Casini for the 18th International Workshop on Non-monotonic Reasoning.

Summary: Propositional KLM-style defeasible reasoning involves a core propositional logic capable of expressing defeasible (or conditional) implications. The semantics for this logic is based on Kripke-like structures known as ranked interpretations. KLM-style defeasible entailment is referred to as rational whenever the defeasible entailment relation under consideration generates a set of defeasible implications all satisfying a set of rationality postulates known as the KLM postulates. In a recent paper Booth et al. proposed PTL, a logic that is more expressive than the core KLM logic. They proved an impossibility result, showing that defeasible entailment for PTL fails to satisfy a set of rationality postulates similar in spirit to the KLM postulates. Their interpretation of the impossibility result is that defeasible entailment for PTL need not be unique. In this paper we continue the line of research in which the expressivity of the core KLM logic is extended. We present the logic Boolean KLM (BKLM) in which we allow for disjunctions, conjunctions, and negations, but not nesting, of defeasible implications. Our contribution is twofold. Firstly, we show (perhaps surprisingly) that BKLM is more expressive than PTL. Our proof is based on the fact that BKLM can characterise all single ranked interpretations, whereas PTL cannot. Secondly, given that the PTL impossibility result also applies to BKLM, we adapt the different forms of PTL entailment proposed by Booth et al. to apply to BKLM.

EOSC as a game-changer in the Social Sciences and Humanities research activities [14] authored by D. Castelli for the Workshop about Language Resources for the SSH Cloud.

Summary: This paper aims to give some insights on how the European Open Science Cloud (EOSC) will be able to influence the Social Sciences and Humanities (SSH) sector, thus paving the way towards innovation. Points of discussion on how the LRs and RIs community can contribute to the revolution in the practice of research areas are provided.

3.3 Contribution to Books

InfraScience members contributed to the following chapters in books.

Data Processing and Analytics for Data-Centric Sciences [9] authored by L. Candela, G. Coro, L. Lelii, G. Panichi, and P. Pagano for Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges.

Summary: The development of data processing and analytics tools is heavily driven by applications, which results in a great variety of software solutions, which often address specific needs. It is difficult to imagine a single solution that is universally suitable for all (or even most) application scenarios and contexts. This chapter describes the data analytics framework that has been designed and developed in the ENVRIplus project to be (a) suitable for serving the needs of researchers in several domains including environmental sciences, (b) open and extensible both with respect to the algorithms and methods it enables and the computing platforms it relies on to execute those algorithms and methods, and (c) open-science-friendly, i.e. it is capable of incorporating every algorithm and method integrated into the data processing framework as well as any computation resulting from the exploitation of integrated algorithms into a “research object” catering for citation, reproducibility, repeatability and provenance.

Virtual Research Environments for Environmental and Earth Sciences: Approaches and Experiences [22] authored by L. Candela for Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges.

Summary: Virtual Research Environments (VREs) are playing an increasingly important role in data centric sciences. Also, the concept is known as Science Gateways in North America where generally the functionality is portal plus workflow deployment and Virtual Laboratories in Australia where the end-user can compose a complete system from the user interface to use of e-Infrastructures by a ‘pick and mix’ process from the offered assets. The key aspect is to provide an environment wherein the end-user - researcher, policymaker, commercial enterprise or citizen scientist - has available with an integrating interface all the assets needed to achieve their objectives. These aspects are explored through different approaches related to ENVRI.

Case Study: ENVRI Science Demonstrators with D4Science [11] authored by L. Candela for Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges.

Summary: Whenever a community of practice starts developing an IT solution for its use case(s) it has to face the issue of carefully selecting “the platform” to use. Such a platform should match the requirements and the overall settings resulting from the specific application context (including legacy technologies and solutions to be integrated and reused, costs of adoption and operation, easiness in acquiring skills and competencies). There is no one-size-fits-all solution that is suitable for all application context, and this is particularly true for scientific communities and their cases because of the wide heterogeneity characterising them. However, there is a large consensus that solutions from scratch are inefficient and services that facilitate the development and maintenance of scientific community-specific solutions do exist. This chapter describes how a set of diverse communities of practice efficiently developed their science demonstrators (on analysing and producing user-defined atmosphere data products, greenhouse gases fluxes, particle formation, mosquito diseases) by leveraging the services offered by the D4Science infrastructure. It shows that the D4Science design decisions aiming at streamlining implementations are effective. The chapter discusses the added value injected in the science demonstrators and resulting from the reuse of D4Science services, especially regarding Open Science practices and overall quality of service.

Learning from the review of Estimating stock status from relative abundance and resilience [20] authored by G. Coro for Marine and Freshwater Miscellanea II, edited by Pauly D., Ruiz-Leotaud V.

Summary: This contribution presents the detailed responses to the peer-review of Froese et al. (2019) "Estimating stock status from relative abundance and resilience" (ICES J. Mar. Sci. 2019) which outlined a method called "AMSY" for inferring biomass trends for stocks for which only catch-per-unit-effort and limited ancillary ('priors') data are available. The responses emphasize that the required priors are legitimate and straightforward to obtain, thus, making AMSY a method of choice in data-sparse situations. This is also a good example of the role of peer-review in validating and improving science.

3.4 Technical Reports

InfraScience members contributed to the following Technical Reports.

Defeasible RDFS via rational closure [13] authored by U. Straccia.

Summary: In the field of non-monotonic logics, the notion of Rational Closure (RC) is acknowledged as a prominent approach. In recent years, RC has gained even more popularity in the context of Description Logics (DLs), the logic underpinning the semantic web standard ontology language OWL 2, whose main ingredients are classes and roles. In this work, we show how to integrate RC within the triple language RDFS, which together with OWL 2 are the two major standard semantic web ontology languages. To do so, we start from ρdf, which is the logic behind RDFS, and then extend it to $\rho df\bot$, allowing to state that two entities are incompatible. Eventually, we propose defeasible $\rho df\bot$ via a typical RC construction. The main features of our approach are: (*i*) unlike most other approaches that add an extra non-monotone rule layer on top of monotone RDFS, defeasible $\rho df\bot$ remains syntactically a triple language and is a simple extension of ρdf by introducing some new predicate symbols with specific semantics. In particular, any RDFS reasoner/store may handle them as ordinary terms if it does not want to take account for the extra semantics of the new predicate symbols; (*ii*) the defeasible ?df? entailment decision procedure is build on top of the $\rho df\bot$ entailment decision procedure, which in turn is an extension of the one for ?df via some additional inference rules favouring an potential implementation; and (*iii*) defeasible $\rho df\bot$ entailment can be decided in polynomial time.

A tale of two 'opens': intersections between Free and Open Source Software and Open Scholarship [33] authored by P. Manghi.

Summary: There is no clear-cut boundary between Free and Open Source Software and Open Scholarship, and the histories, practices, and fundamental principles between the two remain complex. In this study, we critically appraise the intersections and differences between the two movements. Based on our thematic comparison here, we conclude several key things. First, there is substantial scope for new communities of practice to form within scholarly communities that place sharing and collaboration/open participation at their focus. Second, Both the principles and practices of FOSS can be more deeply ingrained within scholarship, asserting a balance between pragmatism and social ideology. Third, at the present, Open Scholarship risks being subverted and compromised by commercial players. Fourth, the shift and acceleration towards a system of Open Scholarship will be greatly enhanced by a concurrent shift in recognising a broader range of practices and outputs beyond traditional peer review and research articles. In order to achieve this, we propose the formulation of a new type of institutional mandate. We believe that there is substantial need for research funders to invest in sustainable open scholarly infrastructure, and the communities that support them, to avoid the capture and enclosure of key research services that would prevent optimal researcher behaviours. Such a shift could ultimately lead to a healthier scientific culture, and a system where competition is replaced by collaboration, resources (including time and people) are shared and acknowledged more efficiently, and the research becomes inherently more rigorous, verified, and reproducible.

EOSC Coordination Day Report [24] authored by E. Lazzeri and D. Castelli.

Summary: This document provides the report for the EOSC Coordination Day that was held in Budapest form 28 to 29 November 2019. Thirty EOSC-related projects come together for two days of discussions on how best they can collaborate in order to ensure a smooth implementation of key aspects of the EOSC. Organised by the EOSCsecretariat.eu project, the meeting saw agreement on a number of actions to bring projects closer together and foster the inclusive and collaborative nature of the EOSC. The meeting also followed up on actions agreed at the first EOSC Concertation meeting organised by European Commission in September 2019.

Turning Open Science and Open Innovation into reality [6] authored by D. Castelli.

Summary: This document summarises the views expressed by the Italian Computing and Data Initiative (ICDI) in response to the open consultation for the EOSC Strategic Research and Innovation Agenda (SRIA), closed on the 31st of August. It provides insightful input and suggestions about the current draft of the SRIA document shared with the wider EOSC community, with the aim of helping to shape the future vision of the European Open Science Cloud.

3.5 Other contributions

InfraScience members contributed to the following publications.

Understanding and managing ocean sustainability: the Blue-Cloud project [10] authored by L. Candela and P. Pagano for ERCIM News - Special theme: Blue Growth.

Summary: The Blue-Cloud flagship project of the Directorate-General (DG) for Research and Innovation Unit of the European Commission is establishing a thematic marine cloud serving the blue economy, marine environment and marine knowledge agendas and the European Open Science Cloud. The project links the horizontal e-infrastructures supported by DG CONNECT and DG GROW, long-term marine data initiatives supported by DG MARE, research infrastructures supported by DG for Research and Innovation and other recently funded thematic clouds.

Predicting the Spread of COVID-19 through Marine Ecological Niche Models [17] authored by G. Coro for ERCIM News - Special theme: Pandemic Modelling and Simulation.

Summary: Researchers from ISTI-CNR (Italy) used marine models, designed to monitor species habitats and invasions, to identify the countries with the highest risk of COVID-19 spread due to climatic and human factors. The model correctly identified most locations where large outbreaks were recorded, independent of population density and dynamics, and is a valuable source of information for smaller-scale population models.

4. Projects

InfraScience was an active member of the consortiums proposing and implementing 11 EU supported projects all focusing on the development of data infrastructures and solutions for various communities of practice. It was also involved in a national project where analytics knowledge and facilities are exploited to provide a quantitative assessment of the effects of the reduced anthropogenic pressure on marine systems during the lockdowns.

ARIADNEplus[2] is a European Union's Horizon 2020 project (grant agreement No. 823914) extending the previous ARIADNE Integrating Activity, which successfully integrated archaeological data infrastructures in Europe, indexing in its registry about 2.000.000 datasets. It extends and supports the research community that the previous project created and further develops the relationships with key stakeholders such as the most important European archaeological associations, researchers, heritage professionals, national heritage agencies and so on. The ARIADNEplus data infrastructure is conceived to offer the availability of Virtual Research Environments where data-based archaeological research may be carried out. The project will furthermore develop a Linked Data approach to data discovery. Innovative services will be made available to users, such as visualization, annotation, text mining and geo-temporal data management. Innovative pilots will be developed to test and demonstrate the innovation potential of the ARIADNEplus approach. Fostering innovation will be a key aspect of the project, with dedicated activities led by the project Innovation Manager. InfraScience is leading two work packages: "Data Integration and Interoperability" to develop, deliver and maintain the ARIADNEplus data and knowledge Cloud and the ARIADNEplus Data Infrastructure; "ARIADNEplus Infrastructure Operation and Management" to (*i*) manage the set of technologies required to operate the ARIADNEplus e-infrastructure, by exploiting the set of services and computational resources provided by the D4Science infrastructure and by supporting the integration of tools, facilities, and services provided by the present project; (*ii*) provide access to the stack of such facilities via Virtual Research Environments, by exploiting the procedures and policies tested and already used by D4Science; (*iii*) manage the software release process covering all stages from integration, through documentation and validation, up to provisioning.

Blue Cloud[3] is a European Union's Horizon 2020 project (grant agreement No. 862409) funded to implement a practical approach to address the potential of cloud based open science to achieve a set of services identifying also longer term challenges to build and demonstrate the Pilot Blue Cloud as a thematic EOSC cloud to support research to better understand and manage the many aspects of ocean sustainability, through a set of five pilot Blue-Cloud demonstrators. It seeks to capitalise on what exists already and to develop and deploy, through a pragmatic workplan, the pilot Blue Cloud as a cyber platform bringing together and providing access to (*i*) multidisciplinary data from observations and models, (*ii*) analytical tools, and (*iii*) computing facilities essential for key blue science use cases. InfraScience is leading the work package "Developing and operating the Blue Cloud VRE, its services and Virtual Labs" called to (*a*) develop and operate the Blue Cloud Virtual Research Environment, (*b*) develop and integrate in the Blue Cloud VRE a data taming service, (*c*) develop and integrate in the Blue Cloud VRE a data analytics service, (*d*) develop and integrate in the Blue Cloud VRE a research object publishing service, (*e*) develop facilities interfacing the Blue Cloud services catalogue with EOSC.

DESIRA[4] is is a European Union's Horizon 2020 project (grant agreement No. 818194) funded to develop a methodology - and a related online tool - to assess the impact of past, current and future digitalization trends of agriculture and rural areas, using the concept of socio-cyber-physical systems – which connect and change data, things, people, plants and animals. Impact analysis will be linked directly to the United

[2] ARIADNEplus Website `ariadne-infrastructure.eu`

[3] Blue Cloud Website `blue-cloud.org`

[4] DESIRA Website `desira2020.eu`

Nation's Sustainable Development Goals. It also contributes to the promotion of the principles of Responsible Research and Innovation. InfraScience is leading the activity "Knowledge Infrastructure: the DESIRA Virtual Research Environment" to design, deliver, and operate the Virtual Research Environment envisaged to serve the needs of the Living Labs. This VRE, a ready-to-use infrastructure for communication exploiting the resources and services operated by D4Science, offers (*i*) a private cloud storage area, equipped with an easy-to-use workspace application designed for use by a wide set of different actors, and the capability to store either private or shared data; (*ii*) social networking applications, where each project member has the possibility to share posts (text, images, and files annotated with hashtags) with VRE members and to collect them in a dedicated News Feed (as in Twitter and Facebook); (*iii*) a private messaging application integrated with the cloud storage to exchange large amount of data securely; (*iv*) an activity tracker and collaborative wiki.

EOSC-Pillar[5] is a European Union's Horizon 2020 project (grant agreement No. 857650) funded to establish an agile and efficient federation model for open science services covering the full spectrum of European research communities by building on representatives of the fast-growing national initiatives for coordinating data infrastructures and services in Italy, France, Germany, Austria and Belgium. It started in July 2019 and will be active up to December 2022. The project aims to contribute to the development of EOSC within a science-driven approach which is efficient, scalable and sustainable and that can be rolled out in other countries. InfraScience is coordinating the contribution of the Italian National Research Council research unit comprising four Institutes: Istituto di Scienza e Tecnologie dell'Informazione A. Faedo (ISTI), Istituto Officina dei Materiali (IOM), Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari (IBIOM), and Istituto di Tecnologie Biomediche (ITB). Moreover, InfraScience is leading the research and development tasks leading to the development of a model and a prototype of a nation service catalog interoperable with EOSC, the development of a catalog driven solution to discover and access the items of a FAIR data space across scattered and heterogeneous data providers, the provisioning of Virtual Research Environments supporting the implementation of case studies in diverse domains.

EOSCsecretariat.eu[6] is a European Union's Horizon 2020 project (grant agreement No. 831644) funded to provide support for the European Open Science Cloud (EOSC), addressing all the specific needs of this digital platform. The project works alongside the community to deliver many of its activities, and it has reserved a substantial portion of its budget for organisations not in the consortium. The EOSCsecretariat.eu's responsibilities include the organisation of EOSC-related events, press and media services, coordination with other related projects, liaison with non-EU countries, efforts to increase pan-European awareness and the provision of a sound legal framework. The project is carried out with the support of highly experienced partners from academia and industry. InfraScience is responsible of the stakeholders engagement. In this role it coordinated activities across projects involved in the implementation of EOSC and established an exchange dialogue with EOSC potential users, like researchers and industry, to provide requirements and feedback during the EOSC development process. It also supports the project by developing and operating a dedicated gateway making available a set of virtual research environments facilitating the collaboration and communication among the members of the task forces the project set up.

I-GENE[7] is a European Union's Horizon 2020 project (grant agreement No. 862714) proposing a new concept of genome editing based on nanotransducers (NTs), aiming to make previously impracticable applications of genome editing and transcriptional regulation by Cas9 safe. This methodology relies on the laser activation of a NT, which triggers consequently a thermo-switchable DNA double strand break or cleavage. The proposed technology implements a concept of multi-input AND gates, where the output (gene editing) is true only if multiple inputs are true at the same time (e.g., NT activation and recognition of 2 different loci). InfraScience provides the I-GENE community with a dedicated gateway and a series of Virtual Research Environments fostering large-scale collaborations where many potentially geographically distributed co-workers can access and process large amounts of data, also by promoting the public debate to support the design of a new strategy/technology for genome editing, ethically acceptable, sustainable and society desirable.

MOVING[8] is a European Union's Horizon 2020 project (grant agreement No. 862739) building capacities and co-develop policy frameworks across Europe to assess how European mountain areas – playing a central role in the well-being of many highly populated European regions –are being impacted by climate change. It establishes new or up-scaled value chains to boost resilience and sustainability of mountain areas. The first step will be to screen traditional and emerging value chains in all European mountain areas. The next step will involve in-depth assessment of vulnerability and resilience of land use, production systems and value chains in 23 mountain regions. The project will use a virtual research environment to promote online interactions amongst actors and new tools to ensure information is accessible by different audiences. InfraScience support the development and operation of the virtual research environment.

OpenAIRE-Advance[9] is a European Union's Horizon 2020 project (grant agreement No. 777541) funded to continue the mission of OpenAIRE to support the Open Access/Open Data mandates in Europe. By sustaining the current successful infrastructure, comprised of a human network and ro-

[5] EOSC-Pillar Website `www.eosc-pillar.eu`

[6] EOSCsecretariat.eu Website `www.eoscsecretariat.eu`

[7] I-GENE Website `i-geneproject.eu`

[8] MOVING Website `www.moving-h2020.eu`

[9] OpenAIRE-Advance Website `www.openaire.eu/advance/`

bust technical services, it consolidates its achievements while working to shift the momentum among its communities to Open Science, aiming to be a trusted e-Infrastructure within the realms of the European Open Science Cloud. In this next phase, OpenAIRE-Advance strives to empower its National Open Access Desks (NOADs) so they become a pivotal part within their own national data infrastructures, positioning Open Access and Open Science onto national agendas. On the technical level OpenAIRE-Advance focuses on the operation and maintenance of the OpenAIRE services, and radically improves the OpenAIRE services on offer by: (*a*) optimizing their performance and scalability, (*b*) refining their functionality based on end-user feedback, (*c*) repackaging them into products, taking a professional marketing approach with well-defined KPIs, (*d*) consolidating the range of services/products into a common e-Infra catalogue to enable a wider uptake. InfraScience was responsible for (*i*) the technical coordination; (*ii*) the operation of the Italian National Open Access Desk (NOAD)[10], which participated to the Research Data Management Task Force; (*iii*) the product management of OpenAIRE-CONNECT[11], (*iv*) work package "Participatory Scholarly Communication", with pilots and technical collaborations with EOSC-Hub, EGI and research infrastructures (EPOS-IT, DARIAH-IT, ELIXIR-GR), (*v*) work package "Optimization & Upgrade of OpenAIRE Technical Services", devoted to technical improvements, scalability optimisation and development of new features of OpenAIRE products (e.g., extension of the Broker service for aggregators in a pilot with LaReferencia, enabling the realization of country portals in a pilot with the Canadian repository network). InfraScience was also responsible for the operation of workflows for the generation of the OpenAIRE Research Graph[12] and the maintenance of the infrastructure for metadata aggregation and full-text collection. InfraScience contributed to the curation of the OpenAIRE Research Graph, integrated project metadata from funders' databases, and developed services to detect duplicates (de-duplication framework), delete wrong links (blacklisting), curate organisation entities (OpenOrgs)[13].

PerformFISH[14] is a European Union's Horizon 2020 project (grant agreement No. 727610) funded to increase the competitiveness of Mediterranean aquaculture by overcoming biological, technical and operational issues with innovative, cost-effective, integrated solutions, while addressing social and environmental responsibility and contributing to "Blue Growth". It adopts a holistic approach constructed with active industry involvement to ensure that Mediterranean marine fish farming matures into a modern dynamic sector, highly appreciated by consumers and society for providing safe and healthy food with a low ecological footprint, and employment and trade in rural, peripheral regions. The project brings together a representative multi-stakeholder, multi-disciplinary consortium to generate, validate and apply new knowledge in real farming conditions to substantially improve the management and performance of the focal fish species, measured through Key Performance Indicators. At the core of PerformFISH design are, (*a*) a link between consumer demand and product design, complemented with product certification and marketing strategies to drive consumer confidence, and (*b*) the establishment and use of a numerical benchmarking system to cover all aspects of Mediterranean marine fish farming performance. InfraScience is leading the activity "Building a Virtual Research Environment (VRE) to Host and Manage Project Data" to deliver (*i*) a set of VREs offering workspace capabilities for supporting the collection, management and controlled sharing of datasets produced by experiments carried out in WPs 1,2,3,4,6. Data sharing will be enabled either between the members of a VRE or between selected users (e.g. colleagues and companies); (*ii*) a VRE supporting KPI data analysis and benchmarking based on production data collected by private companies and securely managed using advanced cryptography and pseudo-anonymisation techniques; (*iii*) a VRE providing access to aggregated and anonymised data to authorised members only.

RISIS 2[15] is a European Union's Horizon 2020 project (grant agreement No. 824091) funded to develop an e-infrastructure that supports full virtual transnational access by researchers in the field of science, technology and innovation to (*a*) an enlarged set of services aimed at meeting field-specific needs (for exploring open data and supporting researchers' analytical capabilities) and (*b*) a set of datasets. InfraScience contributes to the development of the RISIS 2 infrastructure with its infrastructures supporting Open Science, namely D4Science and OpenAIRE. Specifically, the Open Data Virtual Research Environment, empowered by the D4Science, has been equipped with the capability of bridging the RISIS Core Facility Framework and OpenAIRE. This VRE allows delivering tailored and specific datasets collected by OpenAIRE, and selected to satisfy the needs of the RISIS community, to the RISIS project members and community. The Open Data Virtual Research Environment is part of a wider setting involving and all three infrastructures, namely OpenAIRE, D4Science, and RISIS. Its goal is to enrich the RISIS e-Infrastructure in terms of datasets and tools available for the RISIS Community.

Snapshot[16] is an Italian project funded by CNR, whose aim is to provide a quantitative assessment of the effects of the reduced anthropogenic pressure on marine systems during the lockdowns that responded to the COVID-19 pandemic. The 2020 restrictions generated unprecedented, and partially unexpected, human and marine ecosystem dynamics at various levels besides those related to fisheries. By analysing these dynamics in the Italian marine ecosystems, specific cause-

[10] OpenAIRE NOADs `openaire.eu/noad-activities`

[11] OpenAIRE-CONNECT Website `connect.openaire.eu`

[12] OpenAIRE Research Graph Website `graph.openaire.eu`

[13] OpenOrgs Website `https://openaire.eu/openorgs-the-openaire-service-for-bridging-registries-of[illegible]`

[14] PerformFISH Website `performfish.eu/`

[15] RISIS 2 Website `www.risis2.eu`

[16] Snapshot project `http://snapshot.cnr.it/`

effect relationships can be identified and extended to other world ecosystems. The aim of the project is to measure these relationships and the multiple factors involved – including pollution, the economy, fisheries and ecosystem services – to design novel strategies for a more sustainable future. InfraScience – in collaboration with IRBIM-CNR – coordinates the activities regarding ecological niche modelling, fisheries modelling, and climatic data collection and collation.

SoBigData-PlusPlus[17] is a European Union's Horizon 2020 project (grant agreement No. 871042) funded to develop a distributed, Europe-wide, multidisciplinary research infrastructure. This is coupled with the consolidation of a cross-disciplinary European research community. The project builds upon the EU-funded SoBigData project set out to create a research infrastructure delivering an integrated ecosystem for advanced applications of social data mining and Big Data analytics. SoBigData-PlusPlus strengthen infrastructure tools and services by establishing an open platform for the design and performance of large-scale social mining experiments. It delivers specific tools approaching ethics with value-sensitive design integrating values for privacy protection, transparency, and pluralism. InfraScience contributes with its infrastructures supporting Open Science, namely D4Science and OpenAIRE. Specifically, D4Science not only operates the SoBigData e-infrastructure, it enables virtual access to the integrated resources, including existing and newly collected datasets, tools and methods for mining social data. InfraScience VRE technology supports scientists in benefitting from the integration of the integrated resources and from the access to the computational resources, such as the social mining computational engine and the online coding and workflow design frameworks, needed to process these resources. Within this context OpenAIRE provides the online science monitoring dashboard, which monitors and quantifies the outputs of the SoBigData research infrastructure in the scholarly communication ecosystem. It identifies every research product (publications, datasets, software, and other types) produced thanks to the OpenAIRE Research Graph and acts as a single entry point for users to discover, search, browse, and get access to research products related to the infrastructure hosted in several scholarly communication sources (e.g., repositories, journals, archives).

TAILOR[18] is a European Union's Horizon 2020 project (grant agreement No. 952215) with the purpose of building the capacity of providing the scientific foundations for Trustworthy AI in Europe by developing a network of research excellence centres leveraging and combining learning, optimization and reasoning. InfraScience is leading the Trustworthy AI work package aiming at establishing a continuous interdisciplinary dialogue for investigating the methods and methodologies to design, develop, assess, enhance systems that fully implement Trustworthy AI with the ultimate goal to create AI systems that incorporate trustworthiness by design. This activity is organized along the six dimensions of Trustworthy AI: explainability, safety and robustness, fairness, accountability, privacy, and sustainability. Each task aims at advancing knowledge on a specific dimension and puts it in relationships with foundation themes. The overall mission for Trustworthy AI is to combine the various dimensions in the TAILOR research and innovation roadmap. Moreover, to maximize this overall goal and take advantage of any effort in Europe, TAILOR will also interact and collaborate with the activities related to "AI Ethics and Responsible AI" of the proposal Humane-AI-net and will lead the organization of joint scientific actions.

5. Infrastructures

InfraScience leads the development of two large scale and well known infrastructures supporting Open Science, namely D4Science and OpenAIRE. Moreover, the team actively contributed to the development of the European Open Science Cloud by participating in key projects, initiatives and task forces (cf. Sec. 9).

D4Science[19] [4] is an IT infrastructure specifically conceived to support the development and operation of Virtual Research Environments by the as-a-Service provisioning mode. The underlying distributed computing infrastructure is spread across four main sites, geographically distributed, and managed across different administrative domains. The Pisa site is conceived to be the core element of the D4Science computing infrastructure. It realizes a cloud infrastructure completely based on open source technologies aiming at guaranteeing the dynamic allocation of the hardware resources and high availability of the services. Three sites are operated on GARR premises, i.e., the Italian National Research and Education Network. D4Science-based VREs are web-based, community-oriented, collaborative, user-friendly, open-science--enabler working environments for scientists and practitioners willing to work together to perform a certain (research) task. From the end-user perspective, each VRE manifests in a unifying web application (and a set of Application Programming Interfaces (APIs)) (*a*) comprising several components made available by portlets organized in custom pages and menu items and (*b*) running in a plain web browser. Every component is aiming at providing VRE users with facilities implemented by relying on one or more services possibly provisioned by diverse providers. In fact, every VRE is conceived to play the role of a gateway giving seamless access to the datasets and services of interest for the designated community while hiding the diversities originating from the multiplicity of resource providers. Among the components each VRE offers there are some basic ones enacting VRE users to perform their tasks collaboratively, namely: (*a*) a *workspace* component to organise and share any digital artefact of interest; (*b*) a *social networking* component to communicate with coworkers by posts and replies; (*c*) a *data an-*

[17] SoBigData++ Website sobigdata.eu

[18] TAILOR Website tailor-network.eu

[19] www.d4science.org

alytics platform to share and execute analytics methods; (*d*) a *catalogue* component to document and publish any worth sharing digital artifact. In 2020 its user base reached 13,748 active users (+4121 users wrt 2019). These users executed a total of 79,909 working sessions (circa 6,660 working sessions per month) and a total of 286,443,657 analytics tasks (circa 23.8 million tasks per month).

OpenAIRE[20] is a legal entity composed of 49 institutions working to promote and support a sustainable implementation of Open Access and Open Science policies for reproducible science, transparent assessment and omni-comprehensive evaluation. It supports the implementation and alignment of Open Science policies at the international level by developing and promoting the adoption of global open standards and interoperability guidelines to realize a sustainable, participatory, trusted, scholarly communication ecosystem, open to all relevant stakeholders (e.g., research communities, funders, project coordinators) and capable of engaging society and foster innovation. Thanks to the network of National Open Access Desks (NOADs), OpenAIRE supports the implementation of Open Science at the local and national level, supporting researchers, project coordinators, funders and policy makers with training and support activities. Furthermore, the technical infrastructure materializes the OpenAIRE Research Graph: an open, de-duplicated, participatory metadata research graph of interlinked scientific products (including research literature, datasets, software, and other types of research products like workflows, protocols and methods), with access rights information, linked to funding information, research communities and infrastructures. The graph is materialized by collecting more than 210 millions of metadata records from more than 9,000 scholarly data sources worldwide. In addition to the information collected from trusted scholarly data sources, the graph includes metadata and links that are (*i*) asserted by users of the OpenAIRE portals, and (*ii*) inferred by full-text and metadata mining algorithms. Added-value services are built on top of the graph to offer Open Science services to different stakeholders. In 2020, more than 1K repositories implemented the OpenAIRE guidelines for metadata exchange and registered to use the PROVIDE dashboard, 11 research communities used the CONNECT service to offer a thematic discovery portal to their researchers, an average of 25,000 monthly users visited the OpenAIRE portals, with 11,365 registered users. Of those 1,423 from 69 countries used the Link functionality of the portals to add products and links to the OpenAIRE Research Graph. The Italian National Open Access Desks organised 20 training sessions for research performing organizations and started the development of the Italian portal on Open Science[21].

6. Software

InfraScience leads the development of two large scale software systems going hand in had with the two infrastructures described above.

gCube[22] [3] is an open source software toolkit used for building and operating Hybrid Data Infrastructures (namely D4Science) enabling the dynamic deployment of Virtual Research Environments. It consists of hundreds of web services and software libraries overall offering functions including infrastructure development and operation, science gateways development, VRE creation and management, users management, data management, analytics, and open science support. According to OpenHub[23] (statistics collected in December 2021) this software (*i*) has had 19,678 commits made by 48 contributors representing 1,079,842 lines of code (*ii*) is mostly written in Java with a low number of source code comments (*iii*) has a well established, mature codebase maintained by a large development team with stable Y-O-Y commits (*iv*) took an estimated 299 years of effort (COCOMO model) starting with its first commit in October, 2008 ending with its most recent commit. During 2020, 10 releases of this technology have been released (from gCube 4.19 in February 2020 up to gCube 4.27 in December 2020). All these releases have been exploited to enhance the service offered by the D4Science Infrastructure.

D-Net[24] [26] is a framework toolkit designed to support developers at constructing custom aggregative infrastructures in a cost-effective way. D-Net offers data management services capable of providing access to different kinds of external data sources, storing and processing information objects of any data models, converting them into common formats, and exposing information objects to third-party applications through a number of standard access APIs. Its infrastructure enabling services facilitate the construction of domain-specific aggregative infrastructures by selecting and configuring the needed services and easily combining them to form autonomic data processing workflows. The combination of out-of-the box data management services and tools for assembling them into workflows makes the toolkit an appealing starting platform for developers having to face the realization of aggregative infrastructures. In 2020, D-Net featured 8 installations running aggregation systems for (*a*) National aggregators: CeON in Poland, Recolecta in Spain; (*b*) research networks, associations, and infrastructures: EAGLE (Europeana network of Ancient Greek and Latin Epigraphy), EFG (European Film Gateway), OpenAIRE (Open Access Infrastructure for Research in Europe); (*c*) EC projects: PARTHENOS (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies - EC H2020 project GA 654119), ARIADNEplus (Advanced Research Infrastructure for Archaeological Data Networking in Europe - plus - EC H2020 project GA 823914); (*d*) institutions: ISTI Open Portal.

[20] `www.openaire.eu`

[21] Portale Italiano per la Scienza Aperta `www.open-science.it`

[22] gCube Website `www.gcube-system.org`

[23] gCube on Open Hub https://www.openhub.net/p/gCube

[24] D-Net Website `d-net.research-infrastructures.eu`

7. Organised Events

A. Mannocci and P. Manghi co-organised the 2nd workshop on Reframing Research (RefResh 2020), co-located with the 12th International Conference on Social Informatics, on 6-9 October, 2020. L. Candela was member of the program committee.

A. Mannocci co-organised the Scientific Knowledge Graphs Workshop, co-located with the 24th International Conference on Theory and Practice of Digital Libraries, 25-27 August 2020 - Lyon, France. L. Candela was member of the program committee.

L. Candela and P. Manghi were members of the program committee of the 16th Italian Research Conference on Digital Libraries (IRCDL 2020), 30-31 January 2020 - Bari, Italy.

L. Candela was member of the program committee of the 12th International Workshop on Science Gateways (IWSG 2020), June 10-12, 2020 - Cardiff, UK.

E. Lazzeri and P. Manghi organised a Special Session on Open Science at 28th Symposium on Advanced Database Systems (SEBD 2020), June 21-24, 2020 - Villasimius, Sardinia, Italy.

8. Training Activities

InfraScience members organised several training activities and courses mainly related to Open Science:

- Open Science and Research Data Management at University of Pisa, PhD Transversal activities, January - March 2020;
- Webinar: VQR - "Open Access, come e perché", February 2020;
- Train-the-trainers for Università Bicocca Milano, April 2020;
- Praticare l'Open Science nelle Scienze della Terra e dell'Ambiente, November - December 2020;
- CHIST-ERA Course on Open Science and Research Data Management, December 2020;
- Open Science and Research Data Management at Scuola Normale Superiore, 10-18 December 2020;

9. Working Groups, Task Forces, & Interest Groups

InfraScience members chaired the following Working Groups, Task Forces, and Interest Groups:

- *EOSC Future Research Product Publishing Framework Working Group* (A. Bardi) – a WG to define a Research Publishing framework to simplify the adoption of that practice, by enabling the services of research infrastructures to seamlessly integrate repository deposition workflows in the context of the EOSC.
- *EOSC Glossary Interest Group* (D. Castelli) – an IG called to collaboratively provide contribution and feedback towards the creation and development of the EOSC Glossary;
- *Skills and Training Task Force* (E. Lazzeri) – a TF set up within the collaboration agreement among 7 H2020 projects funded by the INFRAEOSC-05 call to share and coordinate the common efforts in the topics of training and skills.
- *EOSC Service and Research Product Catalogues Interest Group* (A. Mannocci) – an IG called to provide a set of "enabling catalogues" that will facilitate the use and re-use of services in support of data-driven research (e.g., computing, storage, scholarly communication, thematic, etc.), data available in a multitude of sources (e.g., repositories, data archives, software archives, libraries, publishers, etc), and scientific products (e.g., publications, research data, research software, other products).
- *ISTI Open Access* (L. Candela) – a WG called to drive the development of the Institute open access and open science policies and practices;
- *ISTI IT infrastructure (S2I2S)* (F. Debole) – a WG called to drive the development of the Institute IT and services;

InfraScience members contributed to the following Working Groups, Task Forces, and Interest Groups:

- *Gruppo di Lavoro Piano Nazionale per la Scienza Aperta* (D. Castelli) – a WG called to develop the italian national plan for Open Science.
- *Commission expert group on National Points of Reference on Scientific Information* (D. Castelli) – Commission lead by EU CNECT - DG Communications Networks, Content and Technology and EU RTD - DG Research and Innovation of Member States' National Points of Reference (NPRs) whose tasks would be to (*i*) co-ordinate the measures listed in the Recommendation C(2012) 4890 final (relating to open access to publications, open research data, preservation of scientific information, and e-infrastructures); (*ii*) to act as interlocutor with the Commission; and (*iii*) to report on the follow-up of the Recommendation.
- *European Commission Open Science Monitor Advisory Board* (E. Lazzeri) – a committee called to support the development of the EU Open Science Monitoring platform by giving advices and suggestions;
- *EOSC Architecture* (P. Manghi) – a WG called to define the technical framework required to enable and sustain an evolving EOSC federation of systems;

- *EOSC Rules of Participation* (P. Pagano) – a WG called to design the Rules of Participation that shall define the rights, obligations governing EOSC transactions between EOSC users, providers and operators;
- *EOSC Skills & Training* (E. Lazzeri) – a WG called to provide a framework for a sustainable training infrastructure to support EOSC in all its phases and ensure its uptake;
- *EOSC Task Force on Scholarly Infrastructures of Research Software* (L. Candela, P. Manghi) – a TF called to established a set of recommendations to allow EOSC to include software, next to other research outputs like publications and data, in the realm of its research artifacts;
- *GOFAIR Discovery Implementation Network* (A. Bardi) – a GO FAIR consortium called to provide interfaces and other user-facing services for data discovery across disciplines.

10. Conclusion

This report documented the research activity performed by the InfraScience research group of the National Research Council of Italy - Institute of Information Science and Technologies (CNR - ISTI) in 2020.

During 2020 InfraScience members contributed to the publishing of 30 papers, to the research and development activities of 12 research projects (11 funded by EU), to the organization of conferences and training events, to several working groups and task forces.

Moreover, the group led the development of two large scale infrastructures for Open Science, i.e., D4Science and OpenAIRE.

Acknowledgments

InfraScience received funding from the European Union's Horizon 2020 research and innovation programme under ARIADNEplus project (grant agreement No. 823914), Blue Cloud project (grant agreement No. 862409), DESIRA project (grant agreement No. 818194), EOSC-Pillar project (grant agreement No. 857650), EOSCsecretariat.eu project (grant agreement No. 831644), I-GENE project (grant agreement No. 862714), MOVING project (grant agreement No. 862739), OpenAIRE-Advance project (grant agreement No. 777541), PerformFISH project (grant agreement No. 727610), RISIS 2 project (grant agreement No. 824091), SoBigData-PlusPlus (grant agreement No. 871042), TAILOR (grant agreement No. 952215).

References

[1] M. Artini, L. Candela, P. Manghi, and S. Giannini. Reposgate: Open science gateways for institutional repositories. In M. Ceci, S. Ferilli, and A. Poggi, editors, *Digital Libraries: The Era of Big Data and Data Science*, pages 151–162, Cham, 2020. Springer International Publishing.

[2] M. Assante, A. Boizet, L. Candela, D. Castelli, R. Cirillo, G. Coro, E. Fernández, M. Filter, L. Frosini, T. Georgiev, G. Kakaletris, P. Katsivelis, R. Knapen, L. Lelii, R. M. Lokers, F. Mangiacrapa, N. Manouselis, P. Pagano, G. Panichi, L. Penev, and F. Sinibaldi. Realizing virtual research environments for the agri-food community: The aginfra plus experience. *Concurrency and Computation: Practice and Experience*, n/a(n/a):e6087, 2020.

[3] M. Assante, L. Candela, D. Castelli, R. Cirilllo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, G. Panichi, C. Perciante, and F. Sinibaldi. The gcube system: Delivering virtual research environments as-a-service. *Future Generation Computer Systems*, 95(n.a.):445–453, 2019.

[4] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, and F. Sinibaldi. Enacting open science by d4science. *Future Generation Computer Systems*, 101:555–563, 2019.

[5] M. Baglioni, P. Manghi, and A. Mannocci. Context-driven discoverability of research data. In M. Hall, T. Merčun, T. Risse, and F. Duchateau, editors, *Digital Libraries for Open Knowledge*, pages 197–211, Cham, 2020. Springer International Publishing.

[6] S. Bassini, T. Boccali, S. Cacciaguerra, D. Castelli, M. Celino, M. Cocco, S. Di Giorgio, A. Giorgetti, G. Kourousias, M. Locati, D. Lucchesi, S. Migliori, G. Pappalardo, L. Perini, C. Petrillo, R. Pugliese, G. Rossi, F. Ruggieri, R. Smareglia, and F. Tanlongo. Turning open science and open innovation into reality. Commission report, European Commission, 2020.

[7] K. Britz, G. Casini, T. Meyer, K. Moodley, U. Sattler, and I. Varzinczak. Principles of klm-style defeasible description logics. *ACM Trans. Comput. Logic*, 22(1), nov 2020.

[8] P. Calyam, N. Wilkins-Diehr, M. Miller, E. H. Brookes, R. Arora, A. Chourasia, D. M. Jennewein, V. Nandigam, M. Drew LaMar, S. B. Cleveland, G. Newman, S. Wang, I. Zaslavsky, M. A. Cianfrocco, K. Ellett, D. Tarboton, K. G. Jeffery, Z. Zhao, J. González-Aranda, M. J. Perri, G. Tucker, L. Candela, T. Kiss, and S. Gesing. Measuring success for a future vision: Defining impact in science gateways/virtual research environments. *Concurrency and Computation: Practice and Experience*, n/a(n/a), 2020.

[9] L. Candela, G. Coro, L. Lelii, G. Panichi, and P. Pagano. *Data Processing and Analytics for Data-Centric Sci-*

ences, pages 176–191. Springer International Publishing, Cham, 2020.

[10] L. Candela and P. Pagano. Understanding and managing ocean sustainability: The blue-cloud project. *ERCIM News*, (123), 2020.

[11] L. Candela, M. Stocker, I. Häggström, C.-F. Enell, D. Vitale, D. Papale, B. Grenier, Y. Chen, and M. Obst. *Case Study: ENVRI Science Demonstrators with D4Science*, pages 307–323. Springer International Publishing, Cham, 2020.

[12] G. Casini, T. Meyer, and I. Varzinczak. Rational Defeasible Belief Change. In D. Calvanese, E. Erdem, and M. Thielscher, editors, *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, pages 213–222, 9 2020.

[13] G. Casini and U. Straccia. Defeasible RDFS via rational closure. Technical report, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", 2020.

[14] D. Castelli. EOSC as a game-changer in the social sciences and humanities research activities. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 37–38, Marseille, France, May 2020. European Language Resources Association.

[15] G. Coro. A global-scale ecological niche model to predict sars-cov-2 coronavirus infection rate. *Ecological Modelling*, 431:109187, 2020.

[16] G. Coro. Open science and artificial intelligence supporting blue growth. *Environmental Engineering and Management Journal*, 19(10):1645–1926, 2020.

[17] G. Coro. Predicting the spread of covid-19 through marine ecological niche models. *ERCIM News*, (124), 2020.

[18] G. Coro, G. Panichi, P. Pagano, and E. Perrone. Nlphub: An e-infrastructure-based text mining hub. *Concurrency and Computation: Practice and Experience*, 33(5):e5986, 2021.

[19] G. Coro and E. Trumpy. Predicting geographical suitability of geothermal power plants. *Journal of Cleaner Production*, 267:121874, 2020.

[20] R. Froese, H. Winker, G. Coro, N. Demirel, A. C. Tsikliras, D. Dimarchopoulou, G. Scarcella, M. L. Deng Palomares, M. Dureuil, and D. Pauly. Learning from the review of estimating stock status from relative abundance and resilience. In D. Pauly and V. Ruiz-Leotaud, editors, *Marine and Freshwater Miscellanea II*, volume 28 of *Fisheries Centre Research Reports*. Institute for the Oceans and Fisheries, The University of British Columbia, Canada, 2020.

[21] I. Huitzil, F. Bobillo, J. Gómez-Romero, and U. Straccia. Fudge: Fuzzy ontology building with consensuated fuzzy datatypes. *Fuzzy Sets and Systems*, 401:91–112, 2020. Fuzzy Measures, Integrals and Quantification in Artificial Intelligence Problems – An Homage to Prof. Miguel Delgado.

[22] K. Jeffery, L. Candela, and H. Glaves. *Virtual Research Environments for Environmental and Earth Sciences: Approaches and Experiences*, pages 272–289. Springer International Publishing, Cham, 2020.

[23] M. J. R. Knapen, R. M. Lokers, L. Candela, and S. Janssen. Aginfra plus: Running crop simulations on the d4science distributed e-infrastructure. In I. N. Athanasiadis, S. P. Frysinger, G. Schimak, and W. J. Knibbe, editors, *Environmental Software Systems. Data Science in Action*, pages 81–89, Cham, 2020. Springer International Publishing.

[24] E. Lazzeri, D. Castelli, L. Marino, S. Garavelli, J. Van Weezel, and N. Ferguson. Eosc coordination day report. Technical Report ISTI-2020-TR/028, ISTI, 2020.

[25] R. M. Lokers, M. J. R. Knapen, L. Candela, S. Hoek, and W. Meijninger. Using virtual research environments in agro-environmental research. In I. N. Athanasiadis, S. P. Frysinger, G. Schimak, and W. J. Knibbe, editors, *Environmental Software Systems. Data Science in Action*, pages 115–121, Cham, 2020. Springer International Publishing.

[26] P. Manghi, M. Artini, C. Atzori, A. Bardi, A. Mannocci, S. La Bruzzo, L. Candela, D. Castelli, and P. Pagano. The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, 48(4):322–354, 2014.

[27] P. Manghi, C. Atzori, M. De Bonis, and A. Bardi. Entity deduplication in big data graphs for scholarly communication. *Data Technol. Appl.*, 54(4):409–435, 2020.

[28] G. Papastefanatos, E. Papadopoulou, M. Meimaris, A. Lempesis, S. Martziou, P. Manghi, and N. Manola. Open science observatory: Monitoring open science in europe. In L. Bellatreche, M. Bieliková, O. Boussaïd, B. Catania, J. Darmont, E. Demidova, F. Duchateau, M. Hall, T. Merčun, B. Novikov, C. Papatheodorou, T. Risse, O. Romero, L. Sautot, G. Talens, R. Wrembel, and M. Žumer, editors, *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, pages 341–346, Cham, 2020. Springer International Publishing.

[29] G. Paterson-Jones, G. Casini, and T. Meyer. BKLM - an expressive logic for defeasible reasoning. In M. V. Martínez and I. Varzinczak, editors, *Proceedings of 18th International Workshop on Non-monotonic Reasoning*, 2020.

[30] A. S. Ribeiro Duarte, C. Liv Nielsen, L. Candela, L. Valentin, F. M. Aarestrup, and H. Vigre. Global food-source identifier (gfi): Collaborative virtual research environment and shared data catalogue for the foodborne outbreak investigation international community. *Food Control*, 121:107623, 2021.

[31] J. Schirrwagen, A. Bardi, A. Czerniak, A. Loehden, N. Rettberg, M. Mertens, and P. Manghi. Data sources and persistent identifiers in the open science research graph of openaire. *International Journal of Digital Curation*, 15(1), 2020.

[32] U. Straccia. How much knowledge is in a knowledge base? introducing knowledge measures (preliminary report). In G. De Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, and J. Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 905–912. IOS Press, 2020.

[33] J. Tennant, R. Agarwal, K. Bazdaric, D. Brassard, T. Crick, D. Dunleavy, T. Evans, N. Gardner, M. Gonzalez-marquez, D. Graziotin, B. Greshake-tzovaras, D. Gunnarsson, J. Havemann, M. Hosseini, D. Katz, M. Knöchelmann, C. Madan, P. Manghi, A. Marocchino, P. Masuzzo, P. Murray-rust, S. Narayanaswamy, G. Nilsonne, J. Pacheco-mendoza, B. Penders, O. Pourret, M. Rera, J. Samuel, T. Steiner, J. Stojanovski, A. Uribe-tirado, R. Vos, S. Worthington, and T. Yarkoni. A tale of two 'opens': intersections between free and open source software and open scholarship. Technical Report 2020/014, ISTI Technical Report, 2020.

InfraScience Members

Michele Artini is a member of the Technical Staff at the Istituto di Scienza e Tecnologie dell'Informazione A. Faedo (ISTI), an institute of the Italian National Research Council (CNR). His skills concern Digital Libraries, e-Infrastructures, Data Management, Web Services, Web Applications and Mobile Applications. Michele joined ISTI in 2005, he worked for several EU Projects such as DELOS, DRIVER, EFG and OpenAIRE, currently, he is working in OpenAIRE-Nexus (EU H2020).

Massimiliano Assante is a researcher of the Istituto di Scienza e Tecnologie dell'Informazione A. Faedo (ISTI), an institute of the Italian National Research Council (CNR). He holds a Ph.D. on Information Engineering and a master degree (M.Sc.) on Information Technologies, both received from the University of Pisa. His research interests include e-Infrastructures, Scientific Repositories, Data Publishing, Virtual Research Environments and NoSQL Data Stores. Massimiliano joined ISTI in 2007, he worked for several EU Projects such as iMarine, EUBrazilOpenBio, D4Science II, D4Science and DILIGENT. Within these projects, he progressively covered different positions, ranging from software engineer (web services and front-end web applications) to analyst, system designer, system integrator, researcher. Currently, he is working in several EU projects (BlueBRIDGE, SoBigData, PARTHENOS, AGINFRA+) and leads the Work Package responsible for Data Access, Discovery, Storage, Analysis and Publishing for the (EU H2020) BlueBRIDGE Project.

Claudio Atzori is a computer science researcher at the National Research Council of Italy, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo". His research activity focuses on digital library management systems, data curation in digital libraries, autonomic service-oriented data infrastructures, and the disambiguation of digital objects in big data graphs. Moreover, he has participated in several EC funded R&D projects: DRIVER-II, EFG, EFG1914, HOPE, EAGLE, OpenAIRE, OpenAIRE-Plus, OpenAIRE2020, OpenAIRE-Advance, OpenAIRE-Connect, OpenAIRE-Nexus, Data4Impact, EOSC-Future as developer, software architect, data analyst, task and work package leader, where his work contributed to the realisation of aggregative data infrastructures for e-science and scholarly communication.

Miriam Baglioni is a (PhD) researcher at InfraScience Laboratory of the Italian National Research Council - Institute of Information Science and Technologies (CNR-ISTI) since 2016. She is currently participating in the EU funded projects OpenAIRE-Nexus, Ariadne Plus and RISIS2. She has worked on Data Mining, Knowledge Discovery, ontologies, social networks and bioinformatics. Her current research interests include data e-infrastructure for science, and science reproducibility.

Alessia Bardi is a PhD researcher in computer science at the Institute of Information Science and Technologies of the Italian National Research Council. She has been involved in EC funded projects for the realisation and operation of aggregative data infrastructures for research communities in the Humanities and Studies of the past (e.g., HOPE - Heritage of the People's Europe, PARTHENOS, Ariadne+) and for the realization of Open Science services like OpenUP, EOSC Future and OpenAIRE projects. In particular, for OpenAIRE she also has the role of product manager for the OpenAIRE CONNECT service. Her research interests include service-oriented architectures, data and metadata interoperability and data infrastructures for e-science and scholarly communication.

Leonardo Candela is computer science senior researcher at the National Research Council of Italy, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo". His research interests are driven by the development of systems and services supporting research infrastructures for science. In particular, he is intertwining virtual research environments, data infrastructures, collaborative working environments, reference models for complex systems, information retrieval, data analytics, data publishing and innovative scholarly communication practices. His research activity is developed by closely connecting research and development. In fact, he has been involved in several EU-funded projects called to develop Digital Libraries & Data Infrastructures and he is the Strategy and

Portfolio Manager of the D4Science.org infrastructure.

Giovanni Casini is a researcher at the National Research Council of Italy, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo". His main research topic is Knowledge Representation and Reasoning, with a particular focus on logical formalisms for uncertain reasoning, belief change, and the Semantic Web. Previously he has worked as a researcher at Scuola Normale Superiore, CSIR (South Africa), University of Pretoria (South Africa), and University of Luxembourg (Luxembourg).

Donatella Castelli is Research Director at Istituto di Scienza e Tecnologie dell'Informazione, "A. Faedo" of the National Research Council of Italy where she leads the InfraScience research group. Under her supervision, the InfraScience team coordinated and participated in several EU and nationally funded projects on Digital Libraries and Research Data Infrastructures. In particular, she has been the co-ordinator of the EU projects that have developed the D4Science infrastructure and technical coordinator of those that have developed the OpenAIRE one. She has participated in experts groups dedicated to the shaping of the European Open Science Cloud. She is currently the Italian member of the EU Group of National contact points for scientific Information. Her research interests include open science data infrastructures and open science scientific approaches. She is author of several research papers in these fields.

Roberto Cirillo is researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. His scientific and professional activity involves the research and development on Data Infrastructures. His research interests include e-Infrastructures, Cloud-based technologies, Virtual Research Environments and NoSQL Data Stores. He is currently member of the BlueBRIDGE EU Project. He was involved in various EU-funded projects including iMARINE, EUBrazil-OpenBio, ENVRI, EGI-ENGAGE. In the past, he has been working on Language Technologies.

Giampaolo Coro is a Physicist with a Ph.D. in Computer Science. His research focuses on Artificial Intelligence, Data Mining and e-Infrastructures. Since 2002, he works on machine learning and signal processing with applications to computational biology, brain-computer interfaces, language technologies and cognitive sciences. The aim of his research is the study and experimentation of models and methodologies to process biological data with an Open Science oriented approach. His approach relies on distributed e-Infrastructures and uses parallel and distributed computing via Cloud-based technologies.

Franca Debole is is a researcher at the Institute of Science and Information Technologies "A. Faedo" of the CNR of Pisa. Graduated in Computer Science at the University of Pisa, she received a PhD in Information Engineering. He has participated in international and national research projects in the field of information retrieval, in the creation of content management systems for multimedia digital libraries and in the field of multilingual search engines. Over the years she has been technical director and involved on several European and National project. Her current research activities range from the digital image processing to techniques for image retrieval and automatic annotation tool. Her technical knowledge ranges from design tools stand alone to web programming techniques. She is also head of a group for IT infrastructure at ISTI-CNR.

Andrea Dell'Amico is a member of the Technical Staff at the Istituto di Scienza e Tecnologie dell'Informazione A. Faedo (ISTI), an institute of the Italian National Research Council (CNR). His skills concern systems administration and integration, automation of systems and services provisioning, configuration and maintenance of large compute and storage infrastructures. He manages the computing and storage facilities of the D4Science.org project. Andrea joined ISTI in 2013 and worked on several EU projects such as BlueBRIDGE, OpenAIRE, Parthenos.

Luca Frosini is researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He has relevant expertise in the area of Virtual Research Environments development. He was involved in various EU-funded projects including DILIGENT, D4Science, EAGLE, PARTHENOS, SoBigData and BlueBRIDGE. Currently, he is Taks Leader of Federated Resources Management in BlueBRIDGE Project. His research interests include Data Infrastructures, Virtual Research Environments, Information Systems, Accounting Systems, and Grid and Cloud Computing.

Sandro La Bruzzo is a member of the Technical Staff at the Institute of Information Science and Technologies "Alessandro Faedo" (ISTI). His skills concern Big Data, Data Analytics & Data infrastructure, Data curation, and aggregation. He is the technical manager of Scholexpler Service. Sandro joined ISTI in 2010; he worked for several EU Projects such as EFG, EAGLE, and OpenAIRE. Currently, he is working in OpenAIRE-Nexus (EU H2020).

Emma Lazzeri, PhD, is currently with Consortium GARR, and also an affiliated researcher at the Institute Information Science and Technologies of the Italian National Research Council in Pisa Italy. She is Open Science manager working on defining strategies, tools and in disseminating and training on Open Science. Emma is deputy member of the European Commission expert group on National Points of Reference on Scientific Information nominated by the Italian Ministry of University and Research. She is member of the European Open Science Cloud (EOSC) Association Advisory Group Task Force on Upskilling Countries to engage in EOSC. She was member of the Executive Board Working Group on Training and Skills, providing a framework for a sustainable training infrastructure to support EOSC in all its phases and ensure its uptake. At national level, she coordinates the Task Force that aims at realising the national

Competence Center for Open Science, FAIR data and EOSC, within the ICDI (Italian computing and data Infrastructure). Emma is the training task leader in EOSC-Pillar project, having as main scope to set up an operational framework for supporting the overall governance of the EOSC, including the coordination between relevant national initiatives. She is also working for CNR in EOSCSecretariat.eu, a EU funded project that supports the European Open Science Cloud (EOSC) governance and co-creation, where she is involved in the Stakeholders engagement. She is member of the Open Science Monitor Expert Group of the European Commission and former member of the Advisory Board for Open Science at CHIST-ERA. She is coordinator of the Italian National Open Access Desks (NOADs) of OpenAIRE and of the Italian Research Data Alliance Node. She is member of the Community of Practice for Training Coordinators in Open Science and of the Education and Communication WG of IOSSG, the Italian Open Science Support Group. Her research interests are in Open Science, including policies, best practices, strategies. Emma is involved as speaker in many international and national conferences in the field of Open Science and Scholarly communication. She holds a PhD in Innovative technology - Telecommunications from Scuola Superiore Sant'Anna, Pisa Italy and a MSc and BSc in Telecommunication engineering from Università di Pisa, Italy.

Lucio Lelii is Researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. His scientific and professional activity involves the Research and Development on Data Infrastructures. He is currently member of the BlueBRIDGE EU Project.

Paolo Manghi is a (PhD) Researcher in computer science at Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of the Consiglio Nazionale delle Ricerche (CNR), in Pisa, Italy. He is the Head of the Scholarly Communication Infrastructures Research Group, working on data e-infrastructures for science and scholarly communication infrastructures, with a focus on technologies supporting open science publishing within and across different disciplines, i.e., computational reproducibility and transparent evaluation of science. Since 2009, he acts as Chief Technical Officer (CTO) for the OpenAIRE AMKE no-profit, operating the European e-infrastructure for Open Science Scholarly Communication. He is the Scientific Coordinator of the H2020 project OpenAIRE-Nexus (Jan 2021) and acted as a scientific coordinator, architect, and/or researcher in the H2020 projects OpenAIRE-Nexus (Jan 2021), OpenAIRE-Connect, OpenAIRE-Advance, and OpenAIRE2020. He is/was involved in the construction and operation of services for the European research infrastructures SoBigDataPlus, PARTHENOS, AriadnePlus, RISIS2, and in the European Open Science Cloud projects EOSCpilot, eInfraCentral, EOSC Secretariat, EOSC-Enhance, and EOSC-Future. He is an active member of Research Data Alliance WGs, member of EC projects advisory boards, of the ResearchObject.org, GreyNet, RD-Switchboard initiative, Open Science Monitor WG for the European Commission, EOSC Architecture WG, GO FAIR GO Inter WG, and World Data System ITO Technical Advisory Committee.

Francesco Mangiacrapa is a computer scientist and researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He has background on geospatial data, technologies, models and standard OGC (like WMS, WFS and so on) for spatial data representation and exchange. His scientific and professional activity includes study and research on Virtual Research Environments and Data Infrastructure, Data Publication, GeoSpatial Data and Open Science. Moreover, his work involve design and development of (Web-)GUI based on several framework (like GWT, Material, Bootstrap and so on) to support his research activity and able to improve community collaboration and exchange of scientific data. Currently, he is working in several EU projects (BlueBRIDGE, SoBigData, PARTHENOS, AGINFRA+) and is responsible for: Data Access and Exchange (Workspace Area), Data Catalogue and Publishing (Catalogue Area) of BlueBRIDGE Project.

Andrea Mannocci is a Research Fellow at ISTI-CNR in Italy. He currently works as a data scientist within the framework of the EU project OpenAIRE Nexus. His research interests span from the analysis of enabling services for Open Science, to Science of Science, complex networks and the analysis of research as a global-scale phenomenon inserted in a delicate socioeconomic and geopolitical context. He obtained his Ph.D. degree in Information Engineering from the University of Pisa (Italy) researching on systems for data flow quality monitoring in data infrastructures. He co-organised the international workshop series on Reframing Research (Refresh2018-2020) held at the European Computational Social Science symposium, and at SocInfo 2020 respectively.

Pasquale Pagano is Senior Researcher at CNR-ISTI. He has a strong background and experience on models, methodologies and techniques for the design and development of distributed virtual research environments (VREs) which require the handling of heterogeneous computational and storage resources, provided by Grid and Cloud based e-Infrastructures, and management of heterogeneous data sources. He participated in the design of the most relevant distributed systems and e- Infrastructure enabling middleware developed by ISTI - CNR. He is currently the Technical Director of the D4Science Data Infrastructure, Technical Director of H2020 BlueBRIDGE project and CNR lead person for the EGI-ENGAGE one. In the past, he has been involved in the iMarine, EUBrazilOpenBio, ENVRI, Venus-C, GRDI2020, D4Science-II, D4Science, Diligent, DRIVER, DRIVER II, BELIEF, BELIEF II, Scholnet, Cyclades, and ARCA European projects.

Giancarlo Panichi is a member of the Technical Staff at the Istituto di Scienza e Tecnologie dell'Informazione A. Faedo (ISTI), an institute of the Italian National Research Council (CNR). His skills concern e-Infrastructures, Web Processing Service, Virtual Research Environments, Data Management, Data Analytics, Web Services, Web Applications and Mobile

Applications. Giancarlo joined ISTI in 2013, he worked for several EU Projects such as iMarine, EUBrazilOpenBio and ENVRI, currently, he is working in BlueBRIDGE Project (EU H2020).

Tommaso Piccioli is a member of the Technical Staff at the A. Faedo Institute of Information Science and Technologies (ISTI). He graduated in Computer Science, with knowledge and responsibility in hardware and software infrastructures design and management, from server farm maintenance to networking, data backup, virtualization environments and systems integration. He was involved since 2005 in the technological support to many projects of the research group including DELOS, Diligent, D4SCIENCE and D4SCIENCE II, iMarine, EUBrazilOpenBio, various OpenAIRE projects, EFG, PerformFISH, PARTHENOS, BlueBRIDGE, RISIS 2, SoBigDataPlus, AriadnePlus.

Fabio Sinibaldi is a Researcher at CNR-ISTI. He holds a degree in computer science engineering with specialization in business management technologies received from the University of Pisa. In his research studies he worked on the design and development of distributed environments' services aimed to manage scientific data, with special attention to Ecological Niche Modelling approaches. These studies involved exploitation of federated Grid and Cloud e-Infrastructures along with Digital Libraries oriented workflow analysis and design, leading to the development of D4Science's Spatial Data Infrastructure. He currently works as Spatial Data Infrastructure designer for D4Science Data Infrastructure under H2020 BlueBRIDGE project and as technology integration manager for EGI-ENGAGE one. In the past he has been involved in the iMarine, EAGLE, EUBrazilOpenBio, ENVRI, Venus-C, D4Science-II, D4Science projects.

Umberto Straccia is a research Director at ISTI - CNR (the Istituto di Scienza e di Tecnologie dell'Informazione - ISTI, an Institute of the National Research Council of Italy - CNR). He received a Ph.D. in computer science from the University of Dortmund, Germany. His research interests include logics for Knowledge Representation and Reasoning (Description Logics, Logic Programming, Answer Set Programming), Semantic Web Languages (OWL, RDFS, RuleML), Fuzzy Logic, Machine Learning (Statistical Relational Learning, Ontology-based Machine Learning), their combination and application.