



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



ISTI Annual Reports

InfraScience research activity report 2022

InfraScience lab., CNR-ISTI, Pisa, Italy



InfraScience research activity report 2022

InfraScience lab.

ISTI-AR-2022/004

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 2022 to highlight the major results. In particular, the InfraScience group confronted with research challenges characterising Data Infrastructures, e-Science, 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 2022 InfraScience members contributed to the publishing of several papers, to the research and development activities of 18 research projects (15 funded by EU), to the organization of conferences and training events, to several working groups and task forces.

Infrastructure, Open Science, Intelligent systems.

Citation

InfraScience lab. *InfraScience research activity report 2022*, ISTI Annual Reports 2022/004. DOI: 10.32079/ISTI-AR-2022/004

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>

InfraScience Research Activity Report 2022

Michele Artini^{ID}, Massimiliano Assante^{ID}, Claudio Atzori^{ID}, Miriam Baglioni^{ID}, Alessia Bardi^{ID}, Pasquale Bove^{ID}, Leonardo Candela^{ID}, Giovanni Casini^{ID}, Donatella Castelli^{ID}, Roberto Cirillo^{ID}, Gianpaolo Coro^{ID}, Michele De Bonis^{ID}, Franca Debole^{ID}, Andrea Dell'Amico^{ID}, Luca Frosini^{ID}, Sandro La Bruzzo^{ID}, Lucio Lelli^{ID}, Paolo Manghi^{ID}, Francesco Mangiacrapa^{ID}, Dario Mangione^{ID}, Andrea Mannocci^{ID}, Enrico Ottonello, Pasquale Pagano^{ID}, Giancarlo Panichi^{ID}, Gina Pavone^{ID}, Tommaso Piccioli, Fabio Sinibaldi^{ID}, Umberto Straccia^{ID}, Franco Zoppi^{ID}

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 2022 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 2022 InfraScience members contributed to the publishing of several papers, to the research and development activities of 21 research projects (16 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

This work is under   

Contents

1	Introduction	1
2	Research topics	2
2.1	Data infrastructures	2
2.2	eScience	2
2.3	Intelligent Systems	2
3	Papers	2
3.1	Contributions to Journals	2
3.2	Contributions to Conferences	8
3.3	Books and contributions to Books	13
3.4	Technical Reports	13
4	Projects	15
5	Infrastructures and Services	20
6	Software	22
7	Datasets	23
8	Organised Events	23
9	Training Activities	24
10	Working Groups, Task Forces, & Interest Groups	24
11	Collaborations	25
12	Conclusion	26

1. Introduction

Science is heavily data and compute-intensive, AI-assisted, participatory, and multidisciplinary. Sharing and publishing scientific results are activities subject to profound reconsideration 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, continuous innovation in IT instruments and approaches.

InfraScience¹ 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)² 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.

¹InfraScience website infrascience.isti.cnr.it

²Institute of Information Science and Technologies website www.isti.cnr.it

This report documents the research activity performed by the group in 2022, 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 produced 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 describes the datasets released by InfraScience. Sec. 8 reports on the organised events. Sec. 9 details the training activity performed by InfraScience. Sec. 10 documents the working groups and task forces InfraScience members participate in. Sec. 11 reports the major collaboration agreements established by InfraScience. Finally, Sec. 12 concludes the report and gives prospects on future research activities.

2. Research topics

The research activities conducted by infraScience members revolve around three major topics: Data Infrastructures, eScience, 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 developing 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., [20, 38, 64], the design and development of integration patterns promoting co-creation, e.g., [8], systematic mapping studies on tools and approaches, e.g., [56], the development of deduplication techniques in big scholarly communication graphs, e.g., [36].

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 studies and proposed approaches for several challenges belonging to the domain including the use of large data sets to study the mechanisms underlying the doing of science, e.g., [58, 59], approaches for managing large scholarly knowledge graphs, e.g., [61, 68], approaches open science-friendly and model-driven aim-

ing at studying a certain phenomenon by aggregating and analysing diverse data, e.g., [35, 33, 34].

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 studied and proposed approaches for challenges including the development of formal models solutions for normative changes [60], reasoning related frameworks and strategies [25, 26, 27], semantic technologies based knowledge production approaches for the archaeological domain [66], systematic mapping studies aiming at understanding the state of the art of recommender systems for science [39].

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 12 articles in journals, 18 papers to conferences and workshops, 2 books or chapter in books, and 12 publications including technical reports and other papers.

3.1 Contributions to Journals

InfraScience members contributed to the following papers published in journals.

New trends in scientific knowledge graphs and research impact assessment [46] by Manghi et al. for Quantitative Science Studies.

Summary: This special issue includes 10 contributions, equally balanced between advances on SKGs and research impact assessment. The papers in the first category introduce several innovative knowledge graphs that enrich classic metadata about articles, patents, and software with further information for exploring these documents more efficiently, identifying insights, and creating more comprehensive analyses of research trends. The articles on impact assessment propose new approaches for key challenges in this field, such as modeling the evolution of credit over time, citing data sets, analyzing research trends on social networks, and predicting citation-based popularity. The contributions address a variety of scientific domains, including computer science, phenomenon-oriented studies, opioids, and COVID-19.

COVID-19 lockdowns reveal the resilience of Adriatic Sea fisheries to forced fishing effort reduction [35] by Coro et al. for Scientific Reports.

Summary: The COVID-19 pandemic provides a major opportunity to study fishing effort dynamics and to assess the response of the industry to standard and remedial actions. Knowing a fishing fleet’s capacity to compensate for effort reduction (i.e., its resilience) allows differentiating governmental regulations by fleet, i.e., imposing stronger restrictions on the more resilient and weaker restrictions on the less resilient. In the present research, the response of the main fishing fleets

of the Adriatic Sea to fishing hour reduction from 2015 to 2020 was measured. Fleet activity per gear type was inferred from monthly Automatic Identification System data. Pattern recognition techniques were applied to study the fishing effort trends and barycentres by gear. The beneficial effects of the lockdowns on Adriatic endangered, threatened and protected (ETP) species were also estimated. Finally, fleet effort series were examined through a stock assessment model to demonstrate that every Adriatic fishing fleet generally behaves like a stock subject to significant stress, which was particularly highlighted by the pandemic. Our findings lend support to the notion that the Adriatic fleets can be compared to predators with medium-high resilience and a generally strong impact on ETP species.

Fig. 1 depicts the fishing effort reduction divided by the maximum sustainable reduction reported for all fishing methods as estimated by the AMSY Bayesian model, used to simulate fishing fleet effort as a stock subject to fishing pressure, represented by governmental restrictions.

Virtual research environments co-creation: The D4Science experience [8] by Assante et al. for Concurrency and Computation: Practice and Experience.

Summary: Virtual research environments are systems called to serve the needs of their designated communities of practice. Every community of practice is a group of people dynamically aggregated by the willingness to collaborate to address a given research question. The virtual research environment provides its users with seamless access to the resources of interest (namely, data and services) no matter what and where they are. Developing a virtual research environment thus to guarantee its uptake from the community of practice is a challenging task. In this article, we advocate how the co-creation driven approach promoted by D4Science has proven to be effective. In particular, we present the co-creation options supported, discuss how diverse communities of practice have exploited these options, and give some usage indicators on the created VREs.

Fig. 2 depicts the service-oriented view of the D4Science architecture.

Filling Gaps in Trawl Surveys at Sea through Spatiotemporal and Environmental Modelling [33] by Coro et al. for Frontiers in Marine Science.

Summary: International scientific fishery survey programmes systematically collect samples of target stocks' biomass and abundance and use them as the basis to estimate stock status in the framework of stock assessment models. The research surveys can also inform decision makers about Essential Fish Habitat conservation and help define harvest control rules based on direct observation of biomass at the sea. However, missed survey locations over the survey years are common in long-term programme data. Currently, modelling approaches to filling gaps in spatiotemporal survey data range from quickly applicable solutions to complex modelling. Most models require setting prior statistical assumptions on spatial distributions, assuming short-term temporal dependency

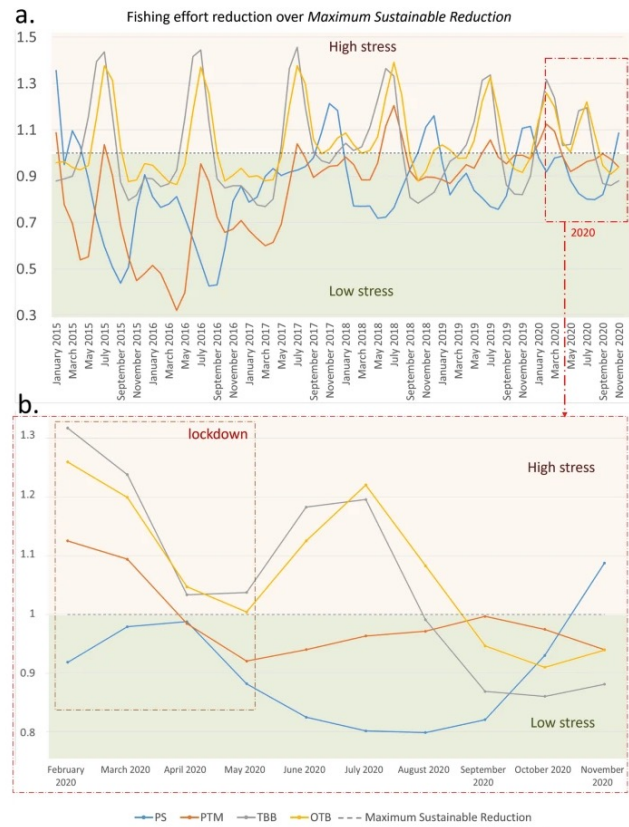


Figure 1. Fishing effort reduction divided by the maximum sustainable reduction reported for all fishing methods as estimated by AMSY: (a) values below the dashed line indicate low fishing hour limitation (low stress), those above the dashed line indicate strong fishing limitation (high stress); (b) detailed representation of the February–November 2020 time series. Acronyms refer to the purse seine (PS), pelagic pair trawl (PTM), beam trawl (TBB) and bottom otter trawl (OTB) fleets. [35]

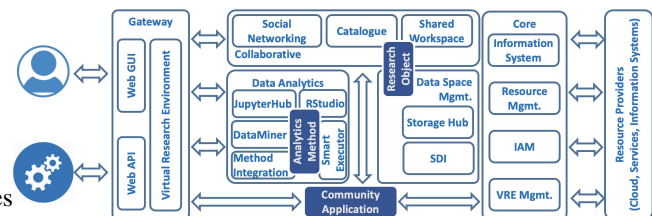


Figure 2. D4Science VRE's: Overall architecture [8]

between the data, and scarcely considering the environmental aspects that might have influenced stock presence in the missed locations. This paper proposes a statistical and machine learning based model to fill spatiotemporal gaps in survey data and produce robust estimates for stock assessment experts, decision makers, and regional fisheries management organizations. We apply our model to the SoleMon survey data in North-Central Adriatic Sea (Mediterranean Sea) for

4 stocks: *Sepia officinalis*, *Solea solea*, *Squilla mantis*, and *Pecten jacobaeus*. We reconstruct the biomass-index (i.e., biomass over the swept area) of 10 locations missed in 2020 (out of the 67 planned) because of several factors, including COVID-19 pandemic related restrictions. We evaluate model performance on 2019 data with respect to an alternative index that assumes biomass proportion consistency over time. Our model's novelty is that it combines three complementary components. A spatial component estimates stock biomass-index in the missed locations in one year, given the surveyed location's biomass-index distribution in the same year. A temporal component forecasts, for each missed survey location, biomass-index given the data history of that haul. An environmental component estimates a biomass-index weighting factor based on the environmental suitability of the haul area to species presence. Combining these components allows understanding the interplay between environmental-change drivers, stock presence, and fisheries. Our model formulation is general enough to be applied to other survey data with lower spatial homogeneity and more temporal gaps than the SoleMon dataset.

Fig. 3 depicts the proposed biomass-index estimation model.

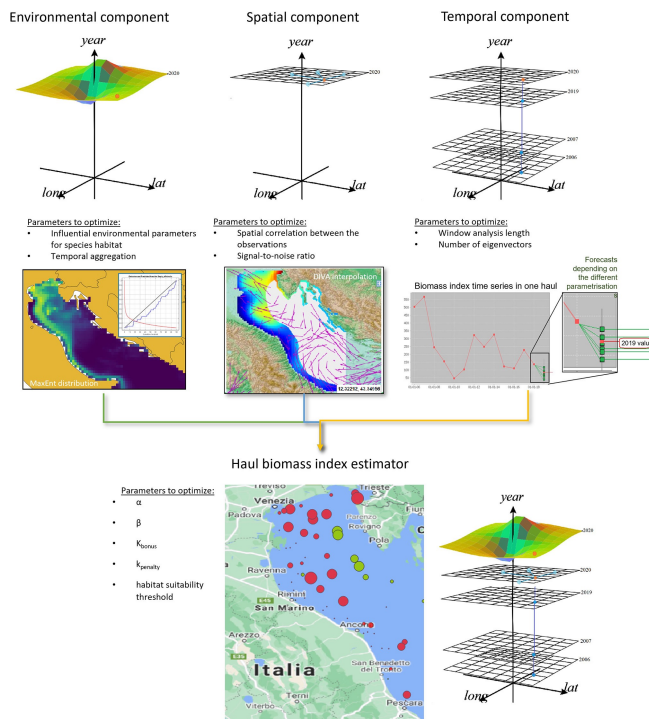


Figure 3. Overview of our overall biomass-index estimation model and its three components, alongside the parameters required by each model [33]

Normative Change: An AGM Approach [60] by Maranhão et al. for the Journal of Applied Logics - IfCoLog Journal.

Summary: Studying normative change is of practical and theoretical interest. Changing legal rules pose interpretation

problems in determining the content of legal rules. The question of interpretation is tightly linked to questions about determining the validity of rules and their ability to produce effects. Different formal models of normative change seem to be better suited to capturing these different dimensions: the dimension of validity appears to be better captured by the AGM approach, while syntactic methods are better suited to modelling how the effects of rules are blocked or enabled. Historically, the AGM approach to belief revision (on which we focus in this article) was the first formal model of normative change. We provide a survey of the AGM approach along with the main criticisms of it. We then turn to a formal analysis of normative change that combines AGM theory and input/output logic, thereby allowing a clear distinction between norms and obligations. Our approach addresses some of the difficulties of normative change, like combining constitutive and regulative rules (and the normative conflicts that may arise from such a combination), revision and contraction of normative systems, as well as contraction of normative systems that combine sets of constitutive and regulative rules. We end our paper by highlighting and discussing some challenges and open problems with the AGM approach regarding normative change

Habitat distribution change of commercial species in the Adriatic Sea during the COVID-19 pandemic [34] by Coro et al. for Ecological Informatics.

Summary: The COVID-19 pandemic has led to reduced anthropogenic pressure on ecosystems in several world areas, but resulting ecosystem responses in these areas have not been investigated. This paper presents an approach to make quick assessments of potential habitat changes in 2020 of eight marine species of commercial importance in the Adriatic Sea. Measurements from floating probes are interpolated through an advection-equation based model. The resulting distributions are then combined with species observations through an ecological niche model to estimate habitat distributions in the past years (2015–2018) at 0.1° spatial resolution. Habitat patterns over 2019 and 2020 are then extracted and explained in terms of specific environmental parameter changes. These changes are finally assessed for their potential dependency on climate change patterns and anthropogenic pressure change due to the pandemic. Our results demonstrate that the combined effect of climate change and the pandemic could have heterogeneous effects on habitat distributions: three species (*Squilla mantis*, *Engraulis encrasicolus*, and *Solea solea*) did not show significant niche distribution change; habitat suitability positively changed for *Sepia officinalis*, but negatively for *Parapenaeus longirostris*, due to increased temperature and decreasing dissolved oxygen (in the Adriatic) generally correlated with climate change; the combination of these trends with an average decrease in chlorophyll, probably due to the pandemic, extended the habitat distributions of *Merluccius merluccius* and *Mullus barbatus* but reduced *Sardina pilchardus* distribution. Although our results are based on approximated data and reliable at a

macroscopic level, we present a very early insight of modifications that will possibly be observed years after the end of the pandemic when complete data will be available. Our approach is entirely based on Findable, Accessible, Interoperable, and Reusable (FAIR) data and is general enough to be used for other species and areas.

Fig. 4 depicts the distribution maps for 2015–2018, 2019, and 2020 for each of the eight analysed species.

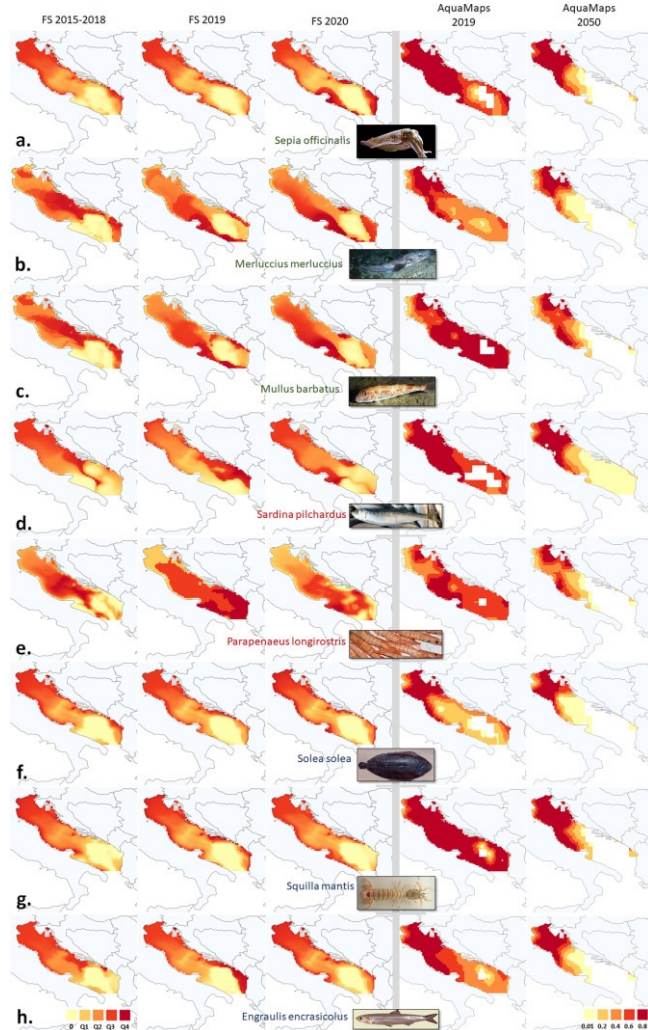


Figure 4. Ecological niches estimated by our floating sensor based (FS) models for 2015–2018, 2019, and 2020, and AquaMaps 2019 and 2020 over the eight analysed species. Coloured species names indicate habitat gain (green), change (red), or stability (blue) in 2020 with respect to 2015–2018. [34]

Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology [32] by Coro et al. for Education and Information Technologies.

Summary: Training through simulation in neonatology relies on sophisticated simulation devices that give realistic

feedback to trainees during simulated scenarios. It aims at training highly specialised medical teams in established operational skills, timely clinical manoeuvres, and successful synergy with other professionals. For effective teaching, it is essential to tailor simulation to trainees' emotional status and communication abilities (human factors), which in turn affect their interaction with the equipment, the environment, and the rest of the team. These factors are crucial to achieving optimal timing and cooperation during a clinical intervention, to the point that they can determine the success of a complex operation such as neonatal resuscitation. Ineffective teams perform in a slow and/or poorly coordinated way and therefore jeopardise positive outcomes. Expert trainers consider human factors as crucial as technical skills. In this context, new technology can help measure learning improvement by quantitatively analysing verbal communication within a medical team. For example, Artificial Intelligence models can work on audio recordings, and draw from extensive historical archives, to extract useful human-factor related information for the trainers. In this study, we present an automatic workflow that supports training through simulation in neonatology by automatically detecting dialogue segments of a simulation session with potentially ineffective communication between team members due to anger, stress, fear, or misunderstandings. Rather than working on audio transcriptions, the workflow analyses syllabic-scale (100–200 ms) spoken dialogue energy and intonation. It uses cluster analysis to identify potentially ineffective communication and extracts the most important related words after audio transcription. Performance is measured against a gold standard containing annotations of 79 minutes of audio recordings from neonatal simulations, in Italian, under different noise conditions (from 4.63 to 14.17 SNR). Our workflow achieves a detection accuracy of 64% and a fair agreement with the gold standard in a challenging context for a speech-processing system, where a commercial automatic speech recogniser reaches just a 9.37% sentence accuracy. The workflow also identifies viable words for trainers to conduct the debriefing session, and can be easily extended to other languages and applications in healthcare. We consider it a promising first step towards introducing new technology to support training through simulation centred on human factors.

Fig. 5 depicts the overall scheme of the proposed workflow.

FDup: a framework for general-purpose and efficient entity deduplication of record collections [36] by De Bonis et al. for PeerJ Computer Science.

Summary: Deduplication is a technique aiming at identifying and resolving duplicate metadata records in a collection. This article describes FDup (Flat Collections Deduper), a general-purpose software framework supporting a complete deduplication workflow to manage big data record collections: metadata record data model definition, identification of candidate duplicates, identification of duplicates. FDup brings two main innovations: first, it delivers a full dedu-

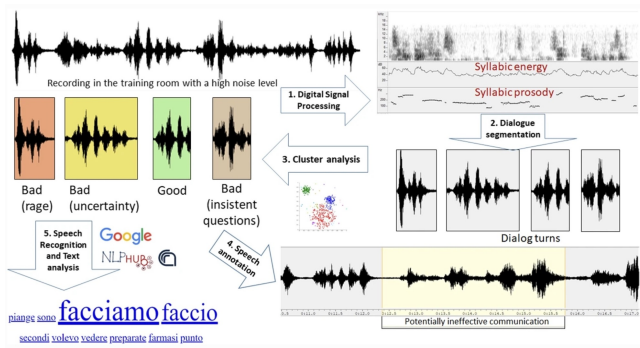


Figure 5. Overall scheme of the proposed workflow: Step 1 (digital signal processing) calculates energy and pitch at a syllabic scale; step 2 (dialogue segmentation) divides the audio into portions with coherent intonation contours (tone units); step 3 (cluster analysis) detects the tone units containing potentially ineffective verbal communication; step 4 (speech annotation) produces an annotation file specifying the intervals of potentially ineffective communication; step 5 (speech recognition and text analysis) transcribes the audio of the tone units through an automatic speech recogniser and extracts a word cloud of ineffective communication keywords. The cloud indicates major communication issues, e.g., “facciamo” (let us do) has a higher weight than “faccio” (I do), suggesting that the team leader is uncertain about which actions to take and would require support to improve leadership and organisation skills. [32]

plication framework in a single easy-to-use software package based on Apache Spark Hadoop framework, where developers can customize the optimal and parallel workflow steps of blocking, sliding windows, and similarity matching function via an intuitive configuration file; second, it introduces a novel approach to improve performance, beyond the known techniques of “blocking” and “sliding window”, by introducing a smart similarity matching function T-match. T-match is engineered as a decision tree that drives the comparisons of the fields of two records as branches of predicates and allows for successful or unsuccessful early-exit strategies. The efficacy of the approach is proved by experiments performed over big data collections of metadata records in the OpenAIRE Research Graph, a known open access knowledge base in Scholarly communication.

Fig. 6 depicts the overall scheme of the proposed workflow.

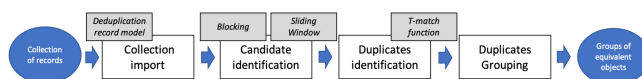


Figure 6. FDup deduplication workflow. [36]

A High-resolution Global-scale Model for COVID-19 Infection Rate [31] by Coro and Bove for ACM Transactions

on Spatial Algorithms and Systems.

Summary: Several models have correlated COVID-19 spread with specific climatic, geophysical, and air pollution conditions, and early models had predicted the lowering of infection cases in Summer 2020. These approaches have been criticized for their coarse assumptions and because they could produce biases if used without considering dynamic factors such as human mobility and interaction. However, human mobility and interaction models alone have not been able to suggest more innovative recommendations than simple social distancing and lockdown, and would definitely need to include information about the base environmental suitability of a World area to COVID-19 spread. This scenario would benefit from a global-scale high-resolution environmental model that could be coupled with dynamic models for large-scale and regional analyses. This article presents a 0.1° high-resolution global-scale probability map of low and high-infection-rates of COVID-19 that uses annual-average surface air temperature, precipitation, and CO₂ as environmental parameters, and Italian provinces as training locations. A risk index calculated on this map correctly identifies 87% of the World countries that reported high infection rates in 2020 and 80% of the low and high infection-rate countries overall. Our model is meant to be used as an additional factor in other models for monthly weather and human mobility. It estimates the base environmental inertia that a geographical place opposes to COVID-19 when mobility restrictions are not in place and can support how much the monthly weather favors or penalizes infection increase. Its high resolution and extent make it consistently usable in global and regional-scale analyses, also thanks to the availability of our results as FAIR data and software as an Open Science-oriented Web service.

Fig. 7 depicts the visual comparison between the high-infection-rate risk map based on a 0.1° resolution MaxEnt model and a previous risk map based on a 0.5° resolution MaxEnt model.

An exploratory approach to archaeological knowledge production [66] by Thanos et al. for International Journal on Digital Libraries.

Summary: The current scientific context is characterized by intensive digitization of the research outcomes and by the creation of data infrastructures for the systematic publication of datasets and data services. Several relationships can exist among these outcomes. Some of them are explicit, e.g., the relationships of spatial or temporal similarity, whereas others are hidden, e.g., the relationship of causality. By materializing these hidden relationships through a linking mechanism, several patterns can be established. These knowledge patterns may lead to the discovery of information previously unknown. A new approach to knowledge production can emerge by following these patterns. This new approach is exploratory because by following these patterns, a researcher can get new insights into a research problem. In the paper, we report our effort to depict this new exploratory approach using Linked Data and Semantic Web technologies (RDF,

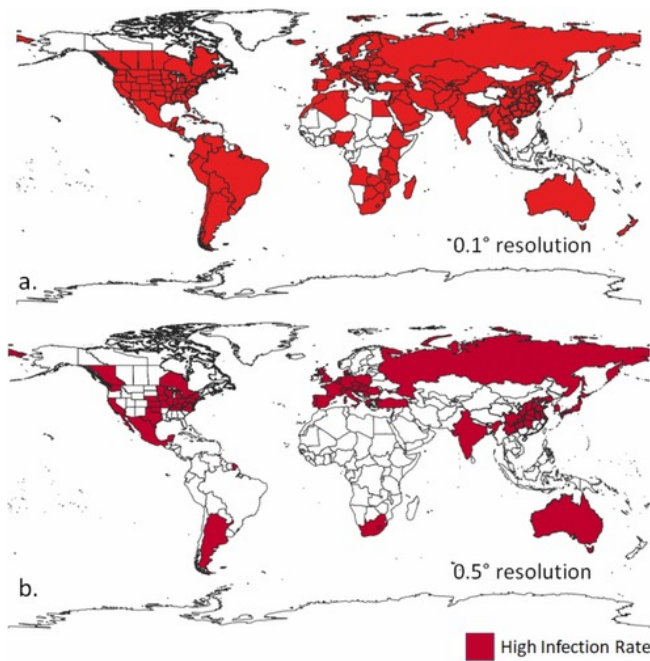


Figure 7. Visual comparison between the high-infection-rate risk map based on a 0.1° resolution MaxEnt model and a previous risk map based on a 0.5° resolution MaxEnt model. [31]

OWL). As a use case, we apply our approach to the archaeological domain.

Fig. 8 depicts the graphical representation of the proposed workflow.

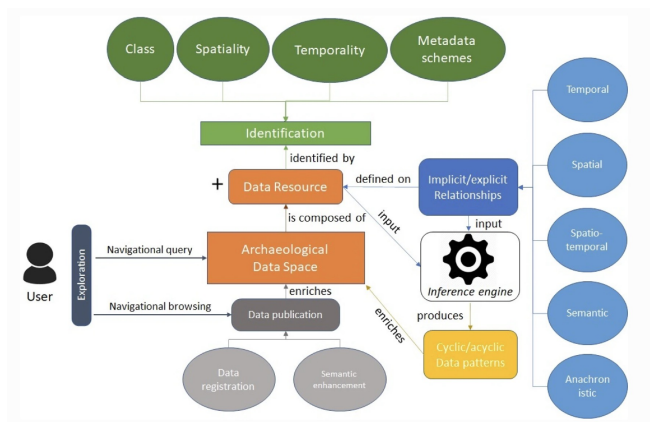


Figure 8. A graphical representation of the proposed workflow that implements an exploratory approach to archaeological knowledge. [66]

The potential effects of COVID-19 lockdown and the following restrictions on the status of eight target stocks in the Adriatic Sea [63] by Scarcella et al. for *Frontiers in Marine Science*.

Summary: The COVID-19 pandemic had major impacts

on the seafood supply chain, also reducing fishing activity. It is worth asking if the fish stocks in the Mediterranean Sea, which in most cases have been in overfishing conditions for many years, may have benefitted from the reduction in the fishing pressure. The present work is the first attempt to make a quantitative evaluation of the fishing effort reduction due to the COVID-19 pandemic and, consequently, its impact on Mediterranean fish stocks, focusing on Adriatic Sea subareas. Eight commercially exploited target stocks (common sole, common cuttlefish, spottail mantis shrimp, European hake, red mullet, anchovy, sardine, and deepwater pink shrimp) were evaluated with a surplus production model, separately fitting the data for each stock until 2019 and until 2020. Results for the 2019 and 2020 models in terms of biomass and fishing mortality were statistically compared with a bootstrap resampling technique to assess their statistical difference. Most of the stocks showed a small but significant improvement in terms of both biomass at sea and reduction in fishing mortality, except cuttlefish and pink shrimp, which showed a reduction in biomass at sea and an increase in fishing mortality (only for common cuttlefish). After reviewing the potential co-occurrence of environmental and management-related factors, we concluded that only in the case of the common sole can an effective biomass improvement related to the pandemic restrictions be detected, because it is the target of the only fishing fleet whose activity remained far lower than expectations for the entire 2020.

Fig. 9 depicts the distribution of the analysed stock (indicated through commonly-accepted acronyms) over regions of fishing sustainability (y-axis) vs regions of population-biomass sustainability (x-axis). The green region includes stocks that are being sustainably fished and in good health (high biomass) status. The red region indicates overfished and depleted stock populations.

NAVIGATOR: an Italian regional imaging biobank to promote precision medicine for oncologic patients [20] by Borgheresi et al. for *European Radiology Experimental*.

Summary: NAVIGATOR is an Italian regional project boosting precision medicine in oncology with the aim of making it more predictive, preventive, and personalised by advancing translational research based on quantitative imaging and integrative omics analyses. The project's goal is to develop an open imaging biobank for the collection and preservation of a large amount of standardised imaging multimodal datasets, including computed tomography, magnetic resonance imaging, and positron emission tomography data, together with the corresponding patient-related and omics-related relevant information extracted from regional healthcare services using an adapted privacy-preserving model. The project is based on an open-source imaging biobank and an open-science oriented virtual research environment (VRE). Available integrative omics and multi-imaging data of three use cases (prostate cancer, rectal cancer, and gastric cancer) will be collected. All data confined in NAVIGATOR (i.e., standard and novel imaging biomarkers, non-imaging data, health agency

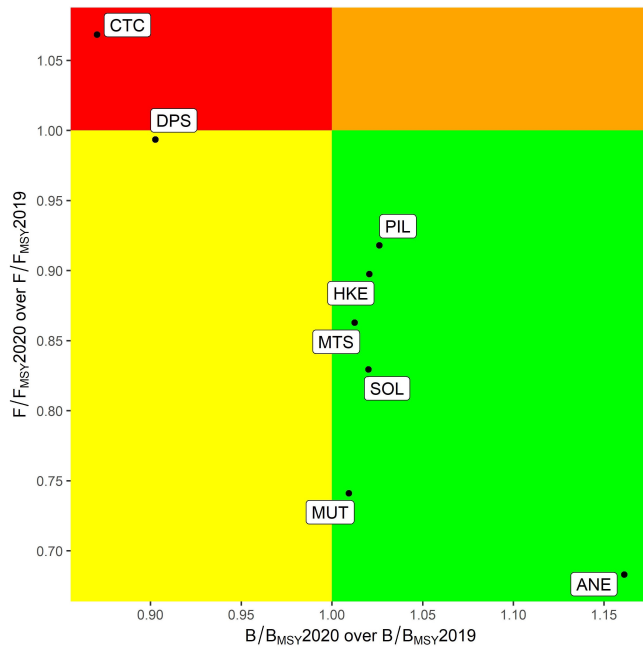


Figure 9. Comparative Kobe plot showing the 2020/2019 ratio of the F/FMSY and B/BMSY outputs from the CMSY/BSM outputs of eight Adriatic Sea Target stocks. [63]

data) will be used to create a digital patient model, to support the reliable prediction of the disease phenotype and risk stratification. The VRE that relies on a well-established infrastructure, called D4Science.org, will further provide a multi-set infrastructure for processing the integrative omics data, extracting specific radiomic signatures, and for identification and testing of novel imaging biomarkers through big data analytics and artificial intelligence.

Fig. 10 depicts the NAVIGATOR virtual research environment (VRE) instantiated over the D4Science platform [7].

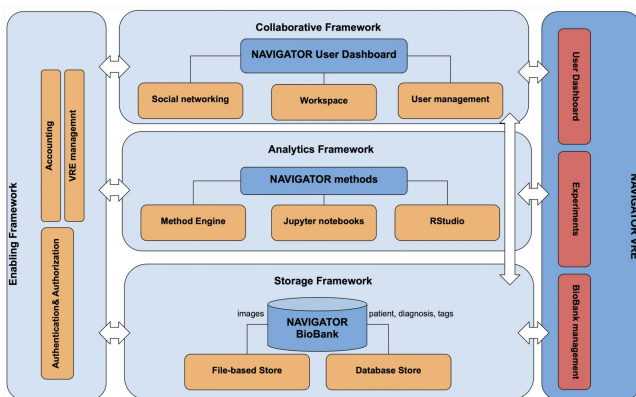


Figure 10. NAVIGATOR virtual research environment (VRE) instantiated over the D4Science platform. [20]

3.2 Contributions to Conferences

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

A Rational Entailment for Expressive Description Logics via Description Logic Programs [27] by Casini and Straccia at Southern African Conference for Artificial Intelligence Research.

Summary: Lehmann and Magidor's rational closure is acknowledged as a landmark in the field of non-monotonic logics and it has also been re-formulated in the context of Description Logics (DLs). We show here how to model a rational form of entailment for expressive DLs, such as *SRQIQ*, providing a novel reasoning procedure that compiles a non-monotone DL knowledge base into a description logic program (dl-program).

Recommender systems for science: a basic taxonomy [39] by Ghannadrad et al. at 18th Italian Research Conference on Digital Libraries (IRCDL 2022).

Summary: The ever-growing availability of research artefacts of potential interest for users calls for helpers to assist their discovery. Artefacts of interest vary for the typology, e.g., papers, datasets, software. User interests are multifaceted and evolving. This paper analyses and classifies studies on recommender systems exploited to suggest research artefacts to researchers regarding the type of algorithm, users and their representations, item typologies and their representation, and evaluation methods used to assess the effectiveness of the recommendations. This study found that most of the current scientific artefacts recommender system focused only on recommending paper to individual researchers, just a few papers focused on dataset recommendation and software recommender system is unprecedented.

Fig. 11 depicts the taxonomy resulting from the systematic mapping study.

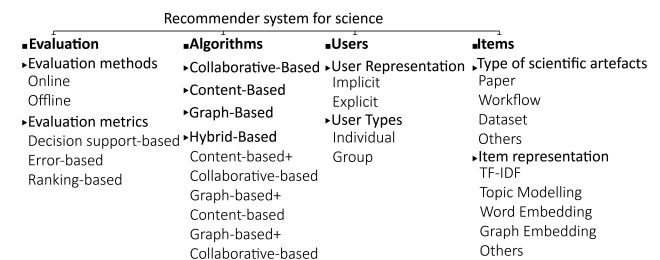


Figure 11. Taxonomy of Recommender system for science. [39]

A taxonomy of tools and approaches for FAIRification [56] by Mangione et al. at 18th Italian Research Conference on Digital Libraries (IRCDL 2022).

Summary: The FAIR principles have drawn a lot of attention since their publication in 2016. A broad range of stakeholders is confronting the implementation of these guiding principles in diverse contexts. This paper identifies and discusses the tools and approaches emerging from stakeholders'

experiences adopting the FAIR principles in practice. In particular, 225 open access grey literature papers (namely, deliverables, milestones and data management plans) on FAIRification have been scrutinised to infer tools and approaches in use. The wealth of emerging tools (477) has been carefully analysed and organised into a comprehensive map highlighting the significant classes of instruments supporting FAIRification. A critical discussion on this collection of tools and approaches and the FAIRification completes the paper.

Fig. 12 depicts the taxonomy resulting from the systematic mapping study by associating the classes with the FAIR principles they contribute to.

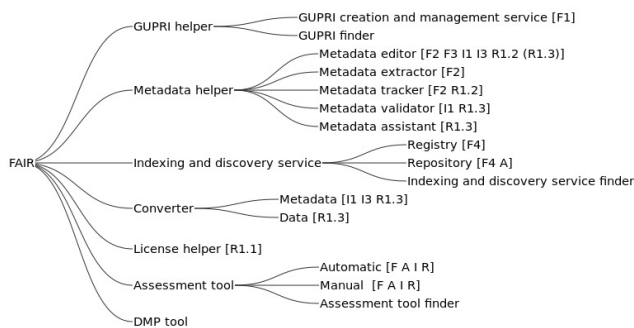


Figure 12. Taxonomy of FAIR tools. [56]

Will open science change authorship for good? Towards a quantitative analysis [58] by Mannocci et al. at 18th Italian Research Conference on Digital Libraries (IRCDL 2022).

Summary: Authorship of scientific articles has profoundly changed from early science until now. If once upon a time a paper was authored by a handful of authors, scientific collaborations are much more prominent on average nowadays. As authorship (and citation) is essentially the primary reward mechanism according to the traditional research evaluation frameworks, it turned to be a rather hot-button topic from which a significant portion of academic disputes stems. However, the novel Open Science practices could be an opportunity to disrupt such dynamics and diversify the credit of the different scientific contributors involved in the diverse phases of the lifecycle of the same research effort. In fact, a paper and research data (or software) contextually published could exhibit different authorship to give credit to the various contributors right where it feels most appropriate. We argue that this can be computationally analysed by taking advantage of the wealth of information in model Open Science Graphs. Such a study can pave the way to understand better the dynamics and patterns of authorship in linked literature, research data and software, and how they evolved over the years.

Towards unsupervised machine learning approaches for knowledge graphs [61] by Minutella et al. at 18th Italian Research Conference on Digital Libraries (IRCDL 2022).

Summary: Nowadays, a lot of data is in the form of Knowledge Graphs aiming at representing information as a set of nodes and relationships between them. This paper proposes an efficient framework to create informative embeddings for node classification on large knowledge graphs. Such embeddings capture how a particular node of the graph interacts with his neighborhood and indicate if it is either isolated or part of a bigger clique. Since a homogeneous graph is necessary to perform this kind of analysis, the framework exploits the metapath approach to split the heterogeneous graph into multiple homogeneous graphs. The proposed pipeline includes an unsupervised attentive neural network to merge different metapaths and produce node embeddings suitable for classification. Preliminary experiments on the IMDb dataset demonstrate the validity of the proposed approach, which can defeat current state-of-the-art unsupervised methods.

Fig. 13 depicts the proposed approach based on a three-step pipeline envisaging the definition of metapaths, the extraction of the embeddings, and the training of the neural network to intelligently aggregate information from different metapaths.



Figure 13. Pipeline proposed in [61] to process knowledge graphs to classifying the nodes given the node attributes and the node neighbourhood.

A preliminary assessment of the article deduplication algorithm used for the OpenAIRE Research Graph [68] by Vichos et al. at 18th Italian Research Conference on Digital Libraries (IRCDL 2022).

Summary: In recent years, a large number of Scholarly Knowledge Graphs (SKGs) have been introduced in the literature. The communities behind these graphs strive to gather, clean, and integrate scholarly metadata from various sources to produce clean and easy-to-process knowledge graphs. In this context, a very important task of the respective cleaning and integration workflows is deduplication. In this paper, we briefly describe and evaluate the accuracy of the deduplication algorithm used for the OpenAIRE Research Graph. Our experiments show that the algorithm has an adequate performance producing a small number of false positives and an even smaller number of false negatives.

Fig. 14 illustrates the proportion of deduplicated entries that have been annotated with each of the classes and summarises the proportion of true and false positives.

A general framework for modelling conditional reasoning - Preliminary report [26] by Casini and Straccia at 19th International Conference on Principles of Knowledge Representation and Reasoning.

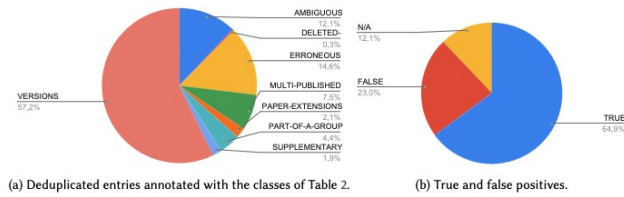


Figure 14. Expert evaluation results in [68]

Summary: We introduce and investigate here a formalisation for conditionals that allows the definition of a broad class of reasoning systems. This framework covers the most popular kinds of conditional reasoning in logic-based KR: the semantics we propose is appropriate for a structural analysis of those conditionals that do not satisfy closure properties associated to classical logics.

A minimal deductive system for RDFS with negative statements [65] by Straccia and Casini at 19th International Conference on Principles of Knowledge Representation and Reasoning.

Summary: The triple language RDFS is designed to represent and reason with *positive* statements only (e.g., “antipyretics are drugs”). In this paper, we extend RDFS to deal with various forms of negative statements under the Open World Assumption (OWA). To do so, we consider pdf , a minimal, but significant RDFS fragment that covers all essential features of RDFS, and then extend it to pdf_{\perp} , allowing express also statements such as “radio therapies are *non drug* treatments”, “Ebola *has no* treatment”, or “opioids and antipyretics are *disjoint* classes”. The main features of our proposal are: (i) pdf_{\perp} remains syntactically a triple language by extending pdf with new symbols with specific semantics and there is no need to revert to the reification method to represent negative triples; (ii) the logic is defined in such a way that any RDFS reasoner/store may handle the new predicates as ordinary terms if it does not want to take account of the extra capabilities; (iii) despite negated statements, every pdf_{\perp} knowledge base is satisfiable; (iv) the pdf_{\perp} entailment decision procedure is obtained from pdf via additional inference rules favouring a potential implementation; and (v) deciding entailment in pdf_{\perp} ranges from P to NP.

Situated conditionals - A brief introduction [29] by Casini et al at 20th International Workshop on Non-Monotonic Reasoning, Part of the Federated Logic Conference (FLoC 2022).

Summary: We extend the expressivity of classical conditional reasoning by introducing *situation* as a new parameter. The enriched conditional logic generalises the defeasible conditional setting in the style of Kraus, Lehmann, and Magidor, and allows for a refined semantics that is able to distinguish, for example, between *expectations* and *counterfactuals*. We introduce the language for the enriched logic and define an appropriate semantic framework for it. We analyse which properties generally associated with conditional reasoning are still satisfied by the new semantic framework,

provide a suitable representation result, and define an entailment relation based on Lehmann and Magidor’s generally-accepted notion of Rational Closure.

Defeasible reasoning in RDFS [25] by Casini and Straccia at 20th International Workshop on Non-Monotonic Reasoning, Part of the Federated Logic Conference (FLoC 2022), Haifa, Israel, August 7-9, 2022.

Summary: RDFS (Resource Description Framework Schema) is a main standard semantic web ontology language that consists of triples (s, p, o) (denoting s is related via p with o). The introduction of non-monotonic formalisms in reasoning with ontologies is useful in particular to deal with situations in which some classes are exceptional and do not satisfy some typical properties of their super classes, as illustrated with the following example.

Developing the EOSC-Pillar RDM training and support catalogue [38] by Garcia et al. at 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022).

Summary: Today’s many research infrastructures and European projects offer training catalogues to store and list multiple forms of learning materials. In EOSC-Pillar project we propose a web application catalogue, which consists of training materials as well as day-to-day operational resources with the aim to support data stewards and other RDM (research data management), FAIR data (findable, accessible, interoperable, reusable) and open science actors. In this paper we briefly describe the scope and technical implementation of the EOSC-Pillar RDM Training and Support Catalogue and how we are addressing current challenges such as metadata standards, controlled vocabularies, curation, quality checking and sustainability.

Fig. 15 displays the catalogue user interface.

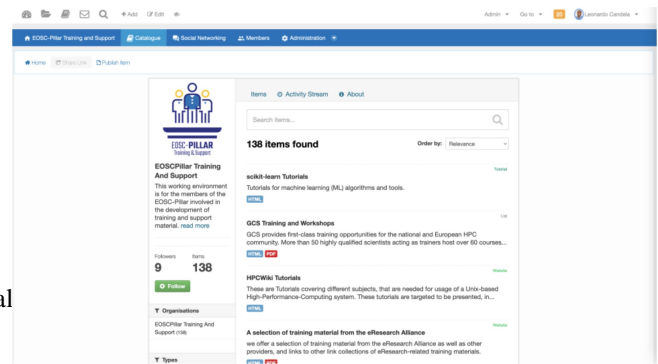


Figure 15. A screenshot of the EOSC-Pillar RDM Training and Support catalogue search interface within the EOSC-Pillar VRE (admin user view). From this page, a catalogue member can browse the catalogue, access the social networking area and other services of the EOSC-Pillar VRE. [38]

Data models for an imaging bio-bank for colorectal, prostate and gastric cancer: the NAVIGATOR project [19] by Berti et al.

at 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).

Summary: Researchers nowadays may take advantage of broad collections of medical data to develop personalized medicine solutions. Imaging bio-banks play a fundamental role, in this regard, by serving as organized repositories of medical images associated with imaging biomarkers. In this context, the NAVIGATOR Project aims to advance colorectal, prostate, and gastric oncology translational research by leveraging quantitative imaging and multi-omics analyses. As Project's core, an imaging bio-bank is being designed and implemented in a web-accessible Virtual Research Environment (VRE). The VRE serves to extract the imaging biomarkers and further process them within prediction algorithms. In our work, we present the realization of the data models for the three cancer use-cases of the Project. First, we carried out an extensive requirements analysis to fulfill the necessities of the clinical partners involved in the Project. Then, we designed three separate data models utilizing entity-relationship diagrams. We found diagrams' modeling for colorectal and prostate cancers to be more straightforward, while gastric cancer required a higher level of complexity. Future developments of this work would include designing a common data model following the Observational Medical Outcomes Partnership Standards. Indeed, a common data model would standardize the logical infrastructure of data models and make the bio-bank easily interoperable with other bio-banks.

Fig. 16 displays one of the Entity-relationship diagrams developed.

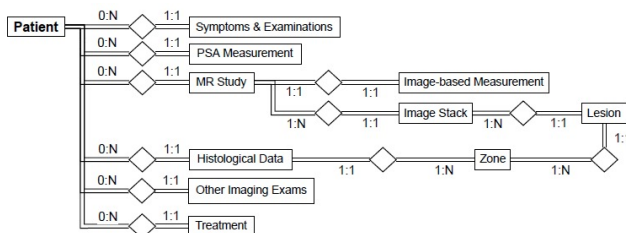


Figure 16. Entity-relationships diagram for the prostate cancer data model. [19]

Blue-Cloud: exploring and demonstrating the potential of Open Science for ocean sustainability [64] by Schaap et al. at 2022 IEEE International Workshop on Metrology for the Sea.

Summary: The Blue-Cloud project is part of 'The Future of Seas and Oceans Flagship Initiative' of the European Commission and runs since October 2019. It has established a pilot cyber platform, providing researchers access to multidisciplinary datasets and derived data products from observations, in-situ and satellite-based, analytical services, and computing facilities essential for blue science to better understand and manage the many aspects of ocean sustainability. A number of core services have been delivered and are now

in a phase of wider dissemination and uptake by marine researchers. Core services are the Federated Data Discovery & Access Service (DD&AS), the Blue-Cloud Virtual Research Environment (VRE), and five Blue-Cloud Virtual Labs.

Fig. 17 displays the Blue-Cloud Virtual Research Environment architecture.

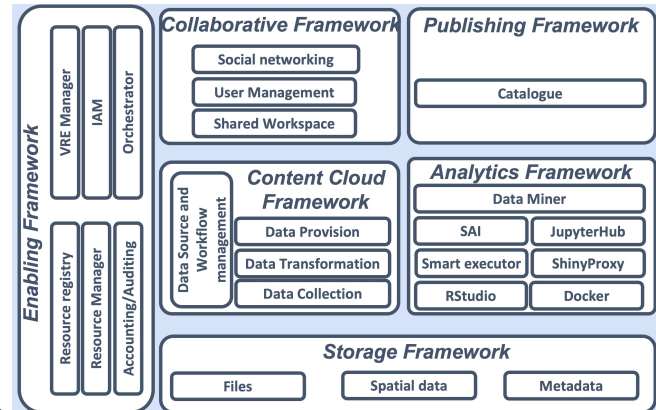


Figure 17. Blue-Cloud Virtual Research Environment architecture [64]

BIP! scholar: a service to facilitate fair researcher assessment [67] by Vergoulis et al. at 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22).

Summary: In recent years, assessing the performance of researchers has become a burden due to the extensive volume of the existing research output. As a result, evaluators often end up relying heavily on a selection of performance indicators like the h-index. However, over-reliance on such indicators may result in reinforcing dubious research practices, while overlooking important aspects of a researcher's career, such as their exact role in the production of particular research works or their contribution to other important types of academic or research activities (e.g., production of datasets, peer reviewing). In response, a number of initiatives that attempt to provide guidelines towards fairer research assessment frameworks have been established. In this work, we present BIP! Scholar, a Web-based service that offers researchers the opportunity to set up profiles that summarise their research careers taking into consideration well-established guidelines for fair research assessment, facilitating the work of evaluators who want to be more compliant with the respective practices.

Fig. 18 displays an indicative profile of a well-known scientist generated by the BIP service.

“Knock Knock! Who’s There?” A study on scholarly repositories’ availability [57] by Mannocci et al. at 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022).

Summary: Scholarly repositories are the cornerstone of modern open science, and their availability is vital for enabling its practices. To this end, scholarly registries such

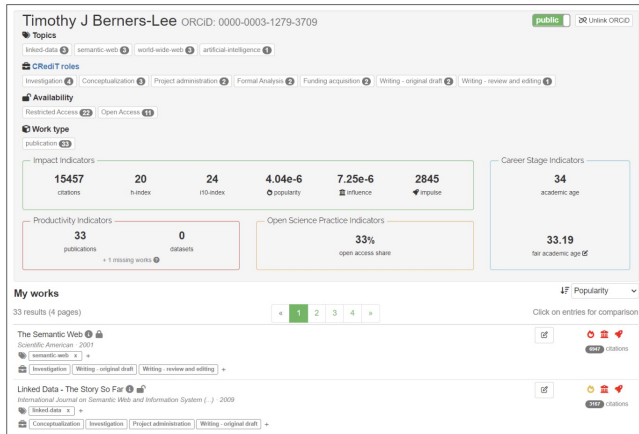


Figure 18. An indicative profile of a well-known scientist. [67]

as FAIRsharing, re3data, OpenDOAR and ROAR give them presence and visibility across different research communities, disciplines, and applications by assigning an identifier and persisting their profiles with summary metadata. Alas, like any other resource available on the Web, scholarly repositories, be they tailored for literature, software or data, are quite dynamic and can be frequently changed, moved, merged or discontinued. Therefore, their references are prone to link rot over time, and their availability often boils down to whether the homepage URLs indicated in authoritative repository profiles within scholarly registries respond or not. For this study, we harvested the content of four prominent scholarly registries and resolved over 13 thousand unique repository URLs. By performing a quantitative analysis on such an extensive collection of repositories, this paper aims to provide a global snapshot of their availability, which bewilderingly is far from granted.

Open Science and authorship of supplementary material. Evidence from a research community [59] by Mannocci et al. at 26th International Conference on Science, Technology and Innovation Indicators (STI 2022).

Summary: Authorship of scientific articles has profoundly changed from early science until now. While once upon a time a paper was authored by a handful of authors, scientific collaborations are much more prominent on average nowadays. As authorship (and citation) is essentially the primary reward mechanism according to the traditional research evaluation frameworks, it turned out to be a rather hot-button topic from which a significant portion of academic disputes stems. However, the novel Open Science practices could be an opportunity to disrupt such dynamics and diversify the credit of the different scientific contributors involved in the diverse phases of the lifecycle of the same research effort. In fact, a paper and research data (or software) contextually published could exhibit different authorship to give credit to the various contributors right where it feels most appropriate. As a preliminary study, in this paper, we leverage the wealth

of information contained in Open Science Graphs, such as OpenAIRE, and conduct a focused analysis on a subset of publications with supplementary material drawn from the European Marine Science (MES) research community. The results are promising and suggest our hypothesis is worth exploring further as we registered in 22% of the cases substantial variations between the authors participating in the publication and the authors participating in the supplementary dataset (or software), thus posing the premises for a longitudinal, large-scale analysis of the phenomenon.

Fig. 19 displays the authorship variation events graph concerning supplementary data.

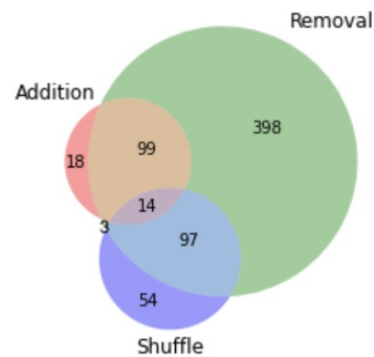


Figure 19. Observer authorship variation events involving supplementary data. [59]

Virtual research environments ethnography: a preliminary study [1] by Arezoumandan et al. at 14th International Workshop on Science Gateways (IWSG 2022).

Summary: Virtual Research Environments, Science Gateways and Virtual Laboratories are systems aiming at serving the needs of their designated communities of practice by providing them with a working environment for performing their tasks. These systems have been proposed and exploited in diverse application domains and scopes ranging from education to simulation, collaboration, and open science. This paper analyses the literature published from 2010 to start characterising this manifold family of systems. In particular, the study identified and analysed a corpus of 1167 research papers to highlight their distribution over time, the most frequent publication venues and the characterising topics.

Figure 20 depicts the distribution of studies by year.

KLM-Style Defeasibility for Restricted First-Order Logic [28] by Casini et al. at 6th International Joint Conference on Rules and Reasoning (RuleML+RR).

Summary: In this paper, we extend the KLM approach to defeasible reasoning beyond the propositional setting. We do so by making it applicable to a restricted version of first-order logic. We describe defeasibility for this logic using a set of rationality postulates, provide a suitable and intuitive semantics for it, and present a representation result characterising the semantic description of defeasibility in terms of

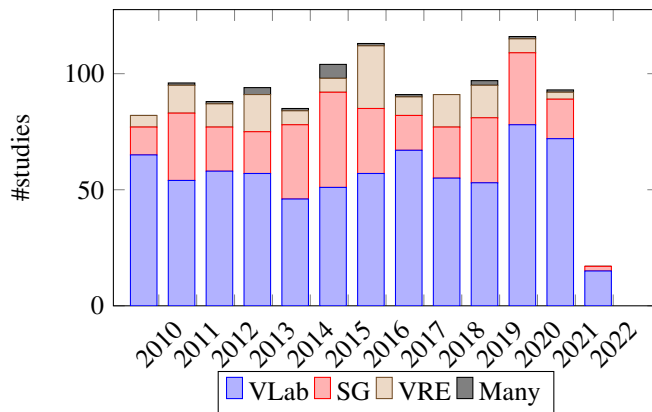


Figure 20. Studies by year.

our postulates. An advantage of our semantics is that it is sufficiently general to be applicable to other restricted versions of first-order logic as well. Based on this theoretical core, we then propose a version of defeasible entailment that is inspired by the well-known notion of Rational Closure as it is defined for defeasible propositional logic and defeasible description logics. We show that this form of defeasible entailment is rational in the sense that it adheres to the full set of rationality postulates.

3.3 Books and contributions to Books

InfraScience members contributed to the following books and chapters in book.

Handbook of Legal AI [30] by Casini et al.

Summary: The Handbook of Legal AI presents a comprehensive overview of the state-of-the-art and trends in the research field of legal AI. The handbook provides a solid introduction to the essentials of the field for newcomers and a selection of advanced issues as a base for future research directions. As the law gets more complex, conflicting, and ever-changing, more advanced methods, most of them come from the Artificial Intelligence (AI) field, are required for analyzing, representing and reasoning on legal knowledge. The discipline that tackles these challenges is now known as “Legal Artificial Intelligence”. Legal AI is experiencing, in particular, in the latest years growth in activity, also at the industrial level, touching a variety of issues which go from the analysis of the textual content of the law, to reasoning about legal interpretation to ethical issues of AI applications in the legal domain (e.g., the artificial judge). This Handbook presents a collection of chapters which evolves around three main topics, namely norm mining (i.e., how to automatically identify, extract, classify and interlink norms from text), reasoning about norms and regulations (i.e., how to derive new legal knowledge from the existing legal knowledge bases in such a way to address automatic legal decision making), and norm enforcement and compliance (i.e., how to check and ensure the compliance of the systems’ requirements with the regulation).

Persistent identifiers and Grey Literature: A PID Project and GreyNet Use Case [37] by Farace et al. in Managing Grey Literature: Technical Services Perspectives.

Summary: Managing grey literature presents many challenges for libraries, from acquisitions to access. One way libraries can promote access to collected grey literature is through the assignment of persistent identifiers (PIDs) to them.

3.4 Technical Reports

InfraScience members contributed to the following Technical Reports.

Open Science repository platforms [47] by Manghi et al. ISTI Technical Report ISTI-TR-2022/009.

Summary: Institutional and thematic repositories today play a key role in scholarly communication and more broadly in scientific workflows. Many institutions and communities have set the ambitious goal of providing an open access repository for their community of users. However, given the amount of expectations from their users, choosing the right solution is often a non-trivial choice. Some platforms may be served out-of-the-box, to be put in operation after straightforward configurations, but are in general less customizable to adhere to specific functional, non-functional, or contextual needs. Other platforms may be instead extremely customizable and flexible but require skilled personnel for their adaptation and deployment. This report performs an analysis of existing state-of-the-art Open Source repository solutions from the functional, operational, and software perspectives. As a result of the analysis, it will factor out the pros and cons of such solutions and identify typical scenarios of adoption.

Bioschemas data sources aggregation to OpenAIRE Research Graph [62] by Ottonello et al. ISTI Technical Report ISTI-2022-TR/010.

Summary: In this report we propose an extended Hadoop-based aggregator for the harvesting of Bioschemas data sources. In this extended hadoop-based aggregator, the downloaded data will be processed according to the consolidated data flow: the original contents will be mapped onto an internal representation that will make them eligible to be integrated in the OpenAIRE research graph.

Research workflows and Open Science [24] by Candela et al. ISTI Technical Report ISTI-2022-TR/026.

Summary: The open science paradigm is increasingly praised and encouraged for improving efficiency through deduplication of efforts and for ultimately accelerating scientific discoveries. Such shift towards a collaborative and inclusive scientific process implies an alteration of the traditional research workflows to include the different dimensions that characterise the new paradigm, from open access to new assessment metrics. This systematic study analyses the open science research workflows proposed so far, highlighting (i) their distribution over time, (ii) the various means and approaches used for communicating them, (iii) the terminology used for denominating them, and (iv) the scientific domains

where workflows were proposed. Moreover, the workflows were analysed and compared concerning a set of open science aspects deriving from the UNESCO Recommendation on Open Science. Overall, a total of 40 relevant studies were identified and analysed, corresponding to 33 unique workflows. The findings highlight (a) the limited effort spent by the research community to propose and communicate workflows oriented to match open science requirements, and (b) the different nuances of the meaning and understanding of open science and the resulting gap between its theoretical aspects and its practical application to the research processes.

Comparison of federated solutions for distributed infrastructures [22] by Candela et al. ISTI Technical Report ISTI-2022-TR/024.

Summary: Federations are an effective mechanism for jointly contributing to a service offering, integrating resources provided by the different federation members. As the open science and the open innovation paradigms are increasingly pursued at the European level, the European Open Science Cloud (EOSC) has been envisaged as a federation of systems for creating a solution fully supporting the research data life cycle across borders and disciplines. This study analyses the federation approaches and solutions adopted by 11 of the major European federated research infrastructures (CLARIN, D4Science, EGI, ELIXIR, ENVRI, EPOS, Gaia-X, GEANT, GEOSS, OpenAIRE and PaNOSC) selected for their representativity following their service offering type and thematic area of expertise. A classification of federated and federating services is proposed in order to highlight the action areas that are to be considered when pursuing the creation of a federated infrastructure with the objective of enabling open science and open innovation.

Research infrastructures: an Open Science quandary [23] by Candela et al. ISTI Technical Report ISTI-2022-TR/025.

Summary: Open science is a disruptive phenomenon characterised by multiple dimensions encompassing the whole research workflow. The magnitude of the changes entailed by the open science paradigm on the research processes requires the research infrastructures to adapt for addressing the technological needs not only of researchers, but of all the stakeholders involved. This study analyses the open science aspects affecting the research workflow and proposes an open science workflow that is then used as a model for examining how the service offering of 11 of the major European research infrastructures (CLARIN, D4Science, EGI, ELIXIR, ENVRI, EPOS, Gaia-X, GEANT, GEOSS, OpenAIRE and PaNOSC) relates to the open science paradigm. In light of the different aspects that characterise open science, a comparative analysis of the 11 service offering is presented and the gaps highlighted.

Data model description of the OpenAIRE Research Graph [42] by La Bruzzo et al. ISTI Technical Report ISTI-2022-TR/031.

Summary: The OpenAIRE Graph (formerly known as the

OpenAIRE Research Graph) is one of the largest open scholarly record collections worldwide, key to fostering Open Science and establishing its practices in daily research activities. Conceived as a public and transparent good, populated out of data sources trusted by scientists, the Graph aims at bringing discovery, monitoring, and assessment of science back into the hands of the scientific community. Imagine a vast collection of research products all linked together, contextualized, and openly available. For the past years, OpenAIRE has been working to gather this valuable record. It is a massive collection of metadata and links between scientific products such as articles, datasets, software, and other research products, entities like organizations, funders, funding streams, projects, communities, and data sources. This technical Report describes the public data model adopted by the OpenAIRE Graph.

OpenAIRE Research Graph deduplication workflow [43] by La Bruzzo et al. ISTI Technical Report ISTI-2022-TR/032.

Summary: The OpenAIRE aggregation workflow can collect metadata records from different providers about the same scholarly work. Each metadata record can carry different information because, for example, some providers are not aware of links to projects, keywords, or other details. Another typical case is when OpenAIRE collects one metadata record from a repository about a pre-print and another from a journal about the published article. To provide correct statistics, OpenAIRE must identify those cases and “merge” the two metadata records so that the scholarly work is counted only once in the statistics OpenAIRE produces. This technical Report describes the Deduplication workflow and technique adopted to deduplicate the OpenAIRE Graph.

OpenAIRE Research Graph: aggregation workflow [44] by La Bruzzo et al. ISTI Technical Report ISTI-2022-TR/033.

Summary: The OpenAIRE Graph (formerly the OpenAIRE Research Graph) is one of the largest open scholarly record collections worldwide. It is key in fostering Open Science and establishing its practices in daily research activities. Conceived as a public and transparent good, populated out of data sources trusted by scientists, the Graph aims at bringing discovery, monitoring, and assessment of science back into the hands of the scientific community. OpenAIRE collects metadata records from more than 70K scholarly communication sources worldwide, including Open Access institutional repositories, data archives, and journals. All the metadata records (i.e., descriptions of research products) are put together in a data lake with records from Crossref, Unpaywall, ORCID, ROR, and information about projects provided by national and international funders. This technical Report describes the main Aggregation Workflow to orchestrate the data aggregation and the implemented mapping from some of the main datasources into the OpenAIRE research graph data model.

OpenOrgs: a tool for the disambiguation of organizations [5] by Artini et al. ISTI Technical Report ISTI-2022-TR/034.

Summary: Organizations appear all over the Research & Innovation ecosystem in different shapes and formats: the same organization may appear with different metadata fields, different names - e.g., full legal name, short or alternative names, acronym. The ambiguity of organizations results in a huge deficiency in the exchange of information, the findability of research products, the monitoring of activities, and ultimately building a linked open scholarly communication system. OpenOrgs combines an automated process and human curation to compensate for the lack of information available and improve the organization's discoverability.

Scholexplorer activity report 2022 [41] by La Bruzzo and Manghi ISTI Technical Report ISTI-2022-TR/035.

Summary: Scholexplorer is a service that accepts publications-data or data-data links from validated sources, builds a de-duplicated graph and provides access to it. ScholExplorer is an implementation of the Scholix initiative (an RDA and WDS). This document is a report on the Scholexplorer installations operation activity after two years of operation, including a detailed set of indicators.

ISTI Open Portal activity report 2022 [4] by Artini et al. ISTI Technical Report ISTI-2022-TR/036.

Summary: ISTI Open Portal is the gateway to the scientific production of the Institute of Information Science and Technologies. It was designed and developed to promote the dissemination of the institute scientific production and its availability according to open access practices. This brief report documents the activities performed in 2022 and gives usage indicators about the service.

D4Science activity report 2022 [9] by Assante et al. ISTI Technical Report ISTI-2022-TR/037.

Summary: D4Science is an IT infrastructure specifically conceived to support the development and operation of Virtual Research Environments by the as-a-Service provisioning mode. This report documents the activities performed in 2022 to develop this infrastructure and support several projects and exploitations.

4. Projects

InfraScience was an active member of the consortiums proposing and implementing 18 research projects (15 were European Union's supported projects) all focusing on the development of data infrastructures and solutions for various communities of practice.

ARIADNEplus³ is a European Union's Horizon 2020 project (grant agreement No. 823914) started in January 2019 and ended in December 2022. It extends 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. The *InfraScience* team 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 Cloud⁴ is a European Union's Horizon 2020 project (grant agreement No. 862409) started in October 2019 and ending in March 2023. It was 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. The *InfraScience* team 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.

CODECS⁵ is a European Union's Horizon Europe project

³ARIADNEplus website ariadne-infrastructure.eu

⁴Blue Cloud website blue-cloud.org

⁵CODECS website <https://www.horizoncodecs.eu/>

(grant agreement No. 101060179) started in October 2022 and ending in September 2026. CODECS works with farmers to develop user-friendly approaches, methods and tools able to document the co-benefits and costs of technologies applied to real contexts. Specifically, it develops a vision of ‘sustainable digitalisation’. By assessing a full range of social, economic and environmental costs and benefits, the CODECS platform will host search, demonstration and assessment tools. The project applies an action research methodology and tests digital technologies via a demonstration farm network. The *InfraScience* team was responsible for the co-ordination of task 8.3 “Ethics, Open Science, data management and gender perspective”, and for gathering information needed for data management and to establish a proper Data Management Plan. Support was also offered at this stage for management choices of research products in line with Open Science and FAIR principles.

DESIRA⁶ is a European Union’s Horizon 2020 project (grant agreement No. 818194) started in June 2019 and ending in May 2023. It was 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 Nation’s Sustainable Development Goals. It also contributes to the promotion of the principles of Responsible Research and Innovation. The *InfraScience* team 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.

EcoScope⁷ is a European Union’s Horizon 2020 project (grant agreement No. 101000302) started in September 2021 and ending in August 2025. It aims to develop an interoperable platform and a robust decision-making toolbox, available through a single public portal, to promote an efficient, ecosystem-based approach to the management of fisheries. It will be guided by policy makers and scientific advisory bodies, and address ecosystem degradation and the anthropogenic impact that are causing fisheries to be unsustainably

exploited across European Seas. In compliance with the Open Science practices, the EcoScope Platform will organise and homogenise climatic, oceanographic, biogeochemical, biological and fisheries datasets for European Seas to a common standard type and format that will be available to the users through interactive mapping layers. The EcoScope Toolbox, a scoring system linked to the platform, will host ecosystem models, socio-economic indicators, fisheries and ecosystem assessment tools that can be used to examine and develop fisheries management and marine policy scenarios as well as maritime spatial planning simulations. Multi-disciplinary groups of end-users and stakeholders will be involved in the design, development and operation of both the platform and the toolbox. Novel assessment methods for data-poor fisheries, including non-commercial species, as well as for biodiversity and the conservation status of protected megafauna, will be used to assess the status of all ecosystem components across European Seas and test new technologies for evaluating the environmental, anthropogenic and climatic impact on ecosystems and fisheries. A series of sophisticated capacity building tools, such as online courses, documentary films, webinars and games, will be available to stakeholders through the EcoScope Academy. By filling these knowledge gaps and developing new methods and tools, the EcoScope project will provide an effective toolbox to decision makers and end-users that will be adaptive to their capacity, needs and data availability. The toolbox will incorporate methods for dealing with uncertainty and deep uncertainty; thus, it will promote efficient, holistic, sustainable, ecosystem-based fisheries management that will aid towards restoring fisheries sustainability and ensuring balance between food security and healthy seas. The *InfraScience* team manages WP4 by contributing to (i) environmental data production and harmonisation, (ii) data mining of vessel transmitted information, (iii) ecological niche modelling via AI models, (iv) biodiversity monitoring indexes, and (v) ecosystem risk assessment.

EOSC Future⁸ is a European Union’s Horizon 2020 project (grant agreement No. 101017536) started in April 2021 and ending in September 2023. It aims to integrate, consolidate, and connect e-infrastructures, research communities, and initiatives in Open Science to develop further the EOSC Portal, EOSC-Core and EOSCExchange of the European Open Science Cloud (EOSC). EOSC Future will unlock the potential of European research via a vision of Open Science for Society by (i) bringing all major stakeholders in the EOSC ecosystem together under one project umbrella to break the disciplinary and community silos and consolidate key EOSC project outputs, (ii) developing scientific use cases in collaboration with the thematic communities showcasing the benefits and societal value of EOSC for doing excellent and interdisciplinary research, (iii) engaging the wider EOSC community and increasing the visibility of EOSC through communications campaigns, marketing strategies, and physical and online engagement events, and (iv) including the

⁶DESIRA website desira2020.eu

⁷EcoScope website <https://ecoscopium.eu/>

⁸EOSC Future website <https://eoscfuture.eu/>

EOSC community in developing the EOSC Portal (including the long tail of science, public and private sectors, and international partners) via co-creation open calls. The *InfraScience* team is contributing in the WP3 to define the architecture and interoperability guidelines and frameworks and in WP4 to the design and development of the Portal Supply Layer (back-office) adapting the existing OpenAIRE services to the EOSC-Core requirements.

EOSC-Pillar⁹ is a European Union's Horizon 2020 project (grant agreement No. 857650) started in July 2019 and ended in December 2022. It was 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. 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.

FAIRCORE4EOSC¹⁰ is a European Union's Horizon Europe project (grant agreement No. 101057264) started in June 2022 and ending in May 2025. It was funded to develop and introduce new components seamlessly integrated with the existing EOSC-Core services, bridging gaps identified in the EOSC Strategic Research and Innovation Agenda (SRIA). It uses existing technologies and services to develop nine new EOSC-Core components to enable EOSC persistent identifiers, an EOSC research software infrastructure and support for advances in EOSC repositories – all of which are important for the FAIR research life cycle. *InfraScience* is responsible for the Research Discovery Graph component (RDGraph). The RDGraph utilises the content of the EOSC catalogue and expands upon it by incorporating additional entities such as Research Activity Identifiers (RAiDs). It provides a range of advanced functionalities that leverage multiple intelligent community-oriented discovery tools developed by the partners of the FAIRCORE4EOSC project.

FOSSR¹¹ is a project funded by the EU Next Genera-

tion Funding in the context of the National Recovery and Resilience Plan. It was funded to become an Italian Open Science Cloud, along the lines of the European Open Science Cloud project, in which to integrate innovative services developed by the project for data collection, data curation and Fairness, and data analysis on economic and societal change. FOSSR fosters the building of an integrated knowledge sharing platform, a single point of access to all the tools and services made available by the Italian nodes of social science infrastructures: ERIC CESSDA, ERIC SHARE and RISIS adopting the common theme of the development of Open Science in the Italian context with the goal of creating a framework of tools and services for the social science scholar community. *InfraScience* contributes with developing and operating the Virtual Research Environment (VRE) service, which nicely complements the FOSSR e-infrastructure offering by enabling the creation of Virtual Laboratories, where Open Science practices are transparently promoted. The VRE service is operated by D4Science and comprises a set of services conceived as a collaborative research platform to boost the capacity to connect to all targeted audiences and increase substantially the interaction among them.

I-GENE¹² is a European Union's Horizon 2020 project (grant agreement No. 862714) started in November 2019 and ending in October 2023. It proposed 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.

MOVING¹³ is a European Union's Horizon 2020 project (grant agreement No. 862739) started in September 2020 and ending in August 2024. It builds 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 upscaled 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

⁹EOSC-Pillar website www.eosc-pillar.eu

¹⁰FAIRCORE4EOSC website <https://faircore4eosc.eu/>

¹¹FOSSR website <http://www.fossr.eu/>

¹²I-GENE website i-geneproject.eu

¹³MOVING website www.moving-h2020.eu

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* supports the development and operation of the virtual research environment.

NAVIGATOR¹⁴ is a project funded by Regione Toscana, started in October 2020 and ending in October 2023. It is called to set the first, regional Virtual Research Environment (VRE) to advance a personalized vision of the clinical management of malignant, solid tumors. The core component will be a Tuscan Biobank of cancer images and imaging biomarkers, set as a shared infrastructure that will support the discovery, test and proof of new models, biomarkers and predictive methods for a better understanding of cancer biology, risk and care. *InfraScience* is involved as service provider, to deliver the NAVIGATOR VRE via the D4Science infrastructure and enable integration of the biobank with the data analysis tools of the platform. The work is performed in collaboration with the SILab team of ISTI.

OpenAIRE Nexus¹⁵ is a European Union's Horizon 2020 project (grant agreement No. 101017452) started in January 2021 and ending in June 2023. The objective of the project is to assemble a suite of services to support researchers, research communities, research performing organisations, policy makers and SME at the adoption, implementation, and monitoring of Open Science practices. The suite is composed of fourteen services, onboarded to the European Open Science Cloud (EOSC), organised in three portfolios: PUBLISH (Zenodo.org, episciences.org, AMNESIA, Argos), DISCOVER (PROVIDE, EXPLORE, CONNECT), MONITOR (OpenAIRE Research Graph, Research Impact Monitoring, UsageCounts, OpenCitations, Scholixplorer, OpenAPC, Open Science Observatory, OpenAIRE AAI). The project also establishes synergies with INFRAEOSC-07 and INFRAEOSC-03 projects to contribute to the interoperability framework for the EOSC. *InfraScience* leads the technical coordination of the project, is responsible for the integration of the services with the EOSC to provide Virtual Access (WP3) and for the contribution to the EOSC Interoperability framework in collaboration with the INFRAEOSC-07 and INFRAEOSC-03 projects (WP7). The group is responsible for the provision of the following services: OpenAIRE Research Graph, Scholixplorer, Broker service (integrated in PROVIDE) and contributes to the delivery of the Research Impact Monitoring, the Open Science Observatory, PROVIDE, EXPLORE, and CONNECT.

PerformFISH¹⁶ is a European Union's Horizon 2020 project (grant agreement No. 727610) started in May 2017 and ended in April 2022. It was 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 2¹⁷ is a European Union's Horizon 2020 project (grant agreement No. 824091) started in January 2019 and ended in December 2022. It was 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.

SerGenCOVID-19¹⁸ is an Italian project funded by CNR,

¹⁴NAVIGATOR website <http://navigator.med.unipi.it/>

¹⁵OpenAIRE Nexus website <https://www.openaire.eu/openaire-nexus-project>

¹⁶PerformFISH website performfish.eu/

¹⁷RISIS 2 website www.risis2.eu

¹⁸SerGenCovid19 website <https://sergencovid.iit.cnr.it/>

started in January 2021 and ending in December 2024. Its goal was to systematically collect clinical data on Italian population affected by Covid-19 to learn more about the individual response to the virus and develop protocols for the management of future patients. *InfraScience* is responsible for the design and development of an IT platform, with the following main components: (a) store and retrieve surveys and results of serological tests; (b) a web interface for the participants to fill the anamnestic questionnaire elaborated by the experts, and to access the results of their serological test; (c) a working environment for data collection by a Virtual Research Environment.

Skills4EOSC¹⁹ is a European Union's Horizon Europe project (grant agreement No. 101058527) started in September 2022 and ending in August 2025. It was funded to unify the existing training landscape into a common and trusted pan-European ecosystem of Competence Centers on open science and data, to accelerate the upskilling of European researchers and data professionals in the field of FAIR and Open Data, intensive-data science, and scientific data management. *InfraScience* is coordinating the contribution of the Italian National Research Council research unit comprising three Institutes: Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), Istituto di Linguistica Computazionale "A. Zampolli" (ILC), and Istituto di Fisica Applicata "N. Carrara" (IFAC). In addition to that, *InfraScience* is leading the task "Collaboration, Coordination and Sharing Platform" called to select a set of agreed and interoperable IT services (e.g., shared workspace, Virtual Labs, Catalogues, Registries, Repositories, Videoconference services facilitating the collaboration, services for supporting common training activities on specific topics) hosted by the Competence Centers to facilitate the continuous alignment, sharing and reuse of material, methodologies and best practices across Competence Centers.

Snapshot²⁰ is an Italian project funded by CNR, started in September 2020 and ended in December 2022. Its goal was 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-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* is responsible for the design and development of an IT platform, with the following main components: (a) store and retrieve surveys and results of serological tests; (b) a web interface for the participants to fill the anamnestic question-

naire elaborated by the experts, and to access the results of their serological test; (c) a working environment for data collection by a Virtual Research Environment.

SoBigData RI PPP²¹ is a European Union's Horizon Europe project (grant agreement No. 101079043) started in October 2022 and ending in September 2025. It was funded to move the SoBigData RI forward from the simple awareness of ethical and legal challenges in social mining to the development of concrete tools that operationalize ethics with value-sensitive design, incorporating values and norms for privacy protection, fairness, transparency, and pluralism. *InfraScience* is responsible for the SoBigData central hub, the component that oversees the technological, administrative, and governance aspects of the RI. The central hub, located in Italy, coordinates all the national nodes, train the staff who will work in them, and establish the methods to provide cutting-edge dynamic digital assets to remote locations without requiring costly on-site expertise. The SoBigData RI follows the System of Systems model, and the national nodes will join and participate based on these principles: autonomy of constituents (independence and evolution); openness (join and leave; dynamic reconfiguration), and distribution (interdependence and interoperability).

SoBigData-PlusPlus²² is a European Union's Horizon 2020 project (grant agreement No. 871042) started in January 2020 and ending in December 2023. It was 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,

¹⁹Skills4EOSC website <https://www.skills4eosc.eu/>

²⁰Snapshot website <http://snapshot.cnr.it/>

²¹SoBigData website sobigdata.eu

²²SoBigData website sobigdata.eu

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).

TAILOR²³ is a European Union's Horizon 2020 project (grant agreement No. 952215) started in September 2020 and ending in August 2023. Its purpose was to 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 and Services

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. 10).

D4Science²⁴ [7] 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 analytics* platform to share and execute analytics methods; (d) a *catalogue* component to document and publish any worth sharing digital artifact. In 2022 its user base reached 19,733 active users (+3,025 users wrt December 2021) (see Fig. 21). These users executed a total of 94,388 working sessions (circa 7,865 working sessions per month) (see Fig. 22) and a total of 99,419,123 analytics tasks (circa 8 millions tasks per month).

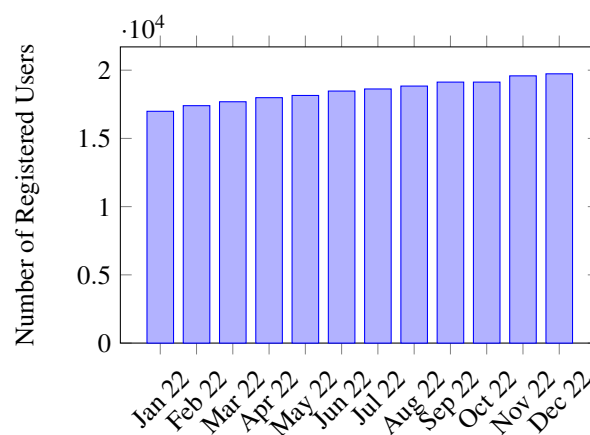


Figure 21. D4Science registered users in 2022.

OpenAIRE²⁵ is the technical infrastructure developed and operated by OpenAIRE AMKE, a legal entity composed of 50 institutions working to promote and support a sustainable implementation of Open Access and Open Science policies for reproducible science, transparent assessment and omnicomprehensive evaluation. OpenAIRE AMKE 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.,

²³TAILOR website tailor-network.eu

²⁴D4Science website www.d4science.org

²⁵OpenAIRE website www.openaire.eu

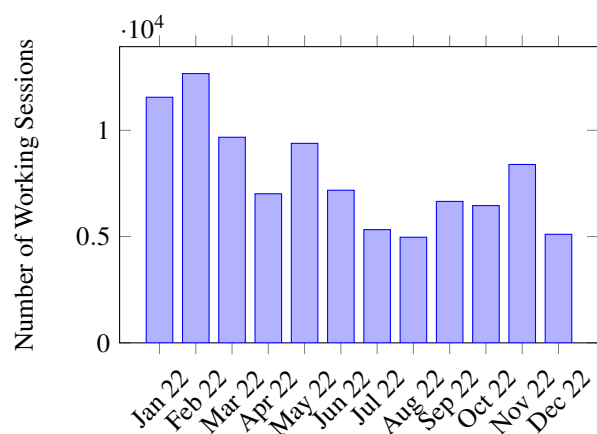


Figure 22. D4Science working sessions in 2022.

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 Graph²⁶: an open, de-duplicated, participatory scientific knowledge 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 materialised by collecting more than 240 millions of metadata records from 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. During 2022, more than 200 repositories implemented the OpenAIRE guidelines for metadata exchange and registered to use the *PROVIDE* dashboard, 28 research communities used the *CONNECT* service to offer a thematic discovery portal to their researchers, and 13 research initiatives used the *CONNECT & MONITOR* services to track their impact; an average of 67,000 monthly users visited the OpenAIRE *EXPLORE* portal, offering search and discovery functionalities over the OpenAIRE Graph. At the end of 2022 the OpenAIRE Graph contained bibliographic records for 148Mi publications, 18Mi datasets, 7Mi other research products and 320K software. As part of the activities of the OpenAIRE National Open Access Desk, support on Open Access and data management is provided on an ongoing basis. In addition, in 2022 several training courses have been provided for different audiences, from undergraduate students and young re-

²⁶The OpenAIRE Graph <https://graph.openaire.eu/>

searchers involved in training networks such as ITN MSCA, to established researchers, including research support staff. The topics covered ranged from Open Science basics and compliance with funding mandates to the Data Management Plan. Besides, OpenAIRE Monitor Institutional Dashboard managers were trained in the use of OpenOrgs [5], an OpenAIRE service to disambiguate organizations. Advocacy and information on Open Science was provided through dedicated talks or venues such as the open-science.it web portal.

InfraScience was also responsible for the development and operation of services for the ISTI community, namely, the *ISTI IT Infrastructure & services* and the *ISTI Open Portal*.

InfraScience was responsible for the management and operation of the *ISTI IT Infrastructure* and its *services* via the S2I2S Working Group (Servizio Infrastruttura Informatica ISTI e Supporto ai Servizi). InfraScience guaranteed the operation of basic services including e-mail, mailing lists, DNS and centralized authentication. In addition to these basic services, the group designed, implemented and made available other research support services (e.g., content collaboration platform, software development service, flexible project management) and provided extensive and timely support concerning their exploitation. Concurrently, the group started the development of the new institute's OpenStack-based IaaS platform, i.e., a modern platform conceived to host the Institute services and facilitate their management in the near future. During 2022 this new infrastructure was equipped with 15 dedicated servers leading to an overall capacity consisting of 840 VCPUs, 5.3TB memory, 250TB disk space. Thus in 2022, the team actually managed two diverse infrastructures, the legacy one to guarantee the operation of the offered services and the new one where the services are likely to be transferred during 2023. The user base of these two infrastructures and its services counted more than 700 active users.

*ISTI Open Portal*²⁷ is a gateway to the scientific production of the Institute of Information Science and Technologies. The gateway is an instance of the RepOSGate technology [3]. It (a) systematically collects the ISTI scientific production from the CNR Institutional Repository, (b) enriches the ISTI products metadata by using information from OpenAIRE, Scihol Explorer [21], and Altmetric²⁸, and (c) make available the open access (self archived) version(s) of ISTI products. In 2022, the gateway had a total of 16,702 item page views and 7971 downloads. Fig. 23 displays the item page views by month. Fig. 24 reports the downloads by month.

During 2022 ISTI supported the creation and operation of the *ISPC Open Portal*²⁹, a gateway replicating the ISTI Open Portal initiative and technology for the needs of the Institute of Heritage Science (Istituto di Scienze del Patrimonio Culturale) of the National Research Council of Italy. The gateway was officially launched in March 2022 and it represents

²⁷ISTI Open Portal <https://openportal.isti.cnr.it>

²⁸Altmetric website <https://www.altmetric.com>

²⁹ISPC Open Portal <https://openportal.ispc.cnr.it/>

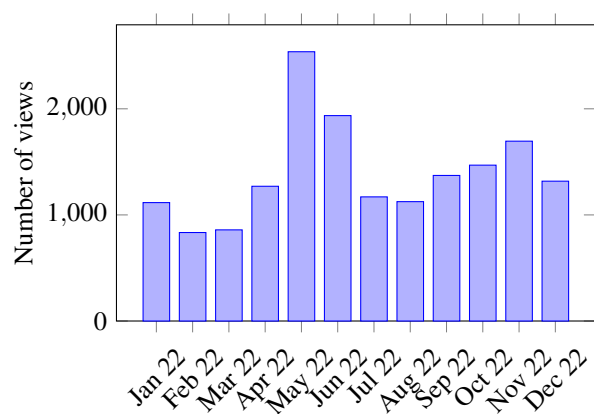


Figure 23. ISTI Open Portal 2022 monthly views.

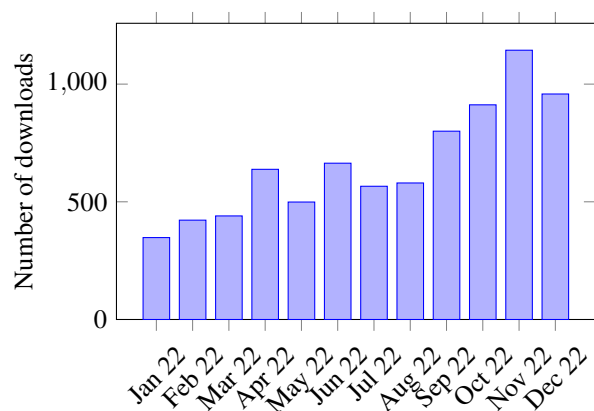


Figure 24. ISTI Open Portal 2022 monthly downloads.

one of the services exploited by CNR-ISPC to implement its open science practices.

InfraScience was responsible for *open-science.it*³⁰, the Italian portal dedicated to Open Science and Open Access. The portal is the result of an initiative developed by the Institute of Information Science and Technologies of the National Research Council of Italy to promote Open Science topics. It originated from the activities of OpenAIRE, the European infrastructure for Open Access, and is supported by the Italian Computing and Data Infrastructure (ICDI) community³¹ comprising stakeholders and experts from 28 Italian Universities and Research Performing Organizations. The portal aims to be a point of reference for the Italian scientific community on issues related to Open Science, Open Access and in general to innovations in academic and scientific communication. The portal was officially launched in December 2021. Since then, the portal has published in-depth articles and news, a special section of FAQs on Open Science and Open Access with a specific legal perspective³², hosts a re-

source catalogue³³ containing more than 50 open science policies of Italian universities and several other documents of interest, and keeps an up-to-date calendar of Open Science events that it makes available for embedding at interested external sites.

6. Software

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

gCube³⁴ [6] 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 OpenHub³⁵ (statistics collected in November 2023) this software (i) has had 28,256 commits made by 53 contributors representing 1,465,563 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 410 years of effort (COCOMO model) starting with its first commit in October, 2008 ending with its most recent commit. During 2022, 16 releases of this technology have been released (from gCube 5.7.0 in January 2022 up to gCube 5.14.0 in December 2022). All these releases have been exploited to enhance the service offered by the D4Science Infrastructure.

D-Net³⁶ [45] is a framework toolkit designed to empower developers in constructing customized aggregative infrastructures in a cost-effective manner. It offers comprehensive data management services that enable access to diverse external data sources, storage and processing of information objects across various data models, conversion into standardized formats, and seamless exposure of information objects to third-party applications through a suite of standardized access APIs. D-Net's infrastructure enabling services streamline the development of domain-specific aggregative infrastructures by facilitating the selection and configuration of required services and enabling their effortless integration into autonomic data processing workflows. This combination of out-of-the-box data management services and workflow assembly tools makes D-Net an attractive starting point for developers tasked with creating aggregative infrastructures. In 2022, D-Net powered nine installations running aggregation systems for: (a) Recolecta, the Spanish National aggregator³⁷; (b) Research networks, associations, and infrastructures – EAGLE (Eu-

³⁰Open-science.it website open-science.it

³¹ICDI website <https://www.icdi.it>

³²<https://open-science.it/faq>

³³<https://open-science.it/catalogue>

³⁴gCube website www.gcube-system.org

³⁵gCube on Open Hub <https://www.openhub.net/p/gCube>

³⁶D-Net website d-net.research-infrastructures.eu

³⁷RECOLECTA website <https://recolecta.fecyt.es>

ropeana network of Ancient Greek and Latin Epigraphy)³⁸, EFG (European Film Gateway)³⁹, OpenAIRE (Open Access Infrastructure for Research in Europe)⁴⁰; (c) EU projects – PARTHENOS (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies - EC H2020 project GA 654119)⁴¹, ARIADNEplus (Advanced Research Infrastructure for Archaeological Data Networking in Europe - plus - EC H2020 project GA 823914)⁴²; (d) Institutions – ISTI Open Portal⁴³, ISCP Open Portal⁴⁴, and INO Open Portal (forthcoming).

7. Datasets

InfraScience released the following datasets.

OpenAIRE Research Graph Dump The dataset includes metadata records of the OpenAIRE graph in json format. Two new releases of the OpenAIRE Research Graph were published [55, 54].

OpenAIRE Research Graph: Dumps for research communities and initiatives. The dataset includes metadata records of research products that are relevant for research communities that have a public OpenAIRE CONNECT Gateway. In particular, three versions were released [51, 52, 53].

OpenAIRE Covid-19 publications, datasets, software and projects metadata. The dataset includes metadata records of publications, research data, software and projects that may be relevant to the Corona Virus Disease (COVID-19) fight. The dump contains records of the OpenAIRE COVID-19 Gateway⁴⁵, identified via full-text mining and inference techniques applied to the OpenAIRE Research Graph. One new version was released [18].

OpenAIRE Research Graph Dump: new collected projects The dataset is updated at every update of the OpenAIRE Graph and contains the new project grants available in the OpenAIRE Graph. One of the main users of the dataset is Zenodo, which uses it to feed the list of grants that users can choose when filling the deposition form. Seven versions were released [11, 12, 13, 14, 15, 16, 17].

OpenAIRE Graph: Dump of funded products The dataset includes metadata records of research products with available funding information. Three new versions were released [49, 50, 48].

³⁸EAGLE website <https://www.eagle-network.eu>

³⁹European Film Gateway website <https://www.europeanfilmgateway.eu>

⁴⁰OpenAIRE Explore website <https://explore.openaire.eu/>

⁴¹PARTHENOS website <https://www.parthenos-project.eu>

⁴²ARIADNEplus website <https://ariadne-infrastructure.eu/>

⁴³ISTI Open Portal <https://openportal.isti.cnr.it>

⁴⁴ISPC Open Portal <https://openportal.ispc.cnr.it/>

⁴⁵OpenAIRE COVID-19 Gateway website <https://covid-19.openaire.eu/>

Books from the OpenAIRE Research Graph This dataset is the subset of the OpenAIRE Graph about research products of type “Book”. One release was published [10].

OpenAIRE ScholXplorer Service: Scholix JSON Dump This dataset contains the dump of links exposed by the OpenAIRE ScholXplorer service in Scholix format [21]. It consists of 417+Mi bi-directional links between literature-dataset and dataset-dataset involving 24+ Mi literature objects and 37+ Mi datasets. A new version was released [40]

8. Organised Events

G. Casini was Program Chair at the *20th International Workshop on Non-Monotonic Reasoning (NMR 2022)*, Part of the Federated Logic Conference (FLoC 2022), Haifa, Israel, August 7-9, 2022. [2] U. Straccia was a member of the Program Committee.

P. Manghi (Program Chair) and L. Candela (Workshop Chair) were members of the Organisation Committee of the *26th International Conference on Theory and Practice of Digital Libraries (TPDL2022)*, Padua, Italy, September 20-23, 2022. D. Castelli was a member of the Senior Program Committee. A. Bardi and A. Mannocci were members of the Program Committee.

P. Manghi and A. Mannocci were co-chairs of the *International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment (Sci-K 2022)*, online, April 26th, 2022. A. Bardi and M. De Bonis were members of the Program Committee.

G. Casini was Doctoral Consortium Chair at the *19th International Conference on Principles of Knowledge Representation and Reasoning (KR 2022)*, part of the Federated Logic Conference (FLoC 2022), Haifa, Israel, July 31 - August 5, 2022.

A. Bardi, L. Candela, P. Manghi, and A. Mannocci were members of the Program Committee of the *18th Italian Research Conference on Digital Libraries (IRCDL 2022)*, Padua, Italy, 24-25 February 2022.

L. Candela was a member of the Program Committee of the *14th International Workshop on Science Gateways (IWSG 2022)*, Trento, Italy, 15th-17th June 2022.

G. Casini and U. Straccia were members of the Program Committee of the *35th International Workshop on Description Logics (DL 2022)*, Haifa, Israel, August 7-10, 2022.

U. Straccia was a member of the Program Committee of the *38th Conference on Uncertainty in Artificial Intelligence (UAI 2022)*, Eindhoven, Netherlands, August 1-5, 2022; the *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2022)*, Milan, Italy, July 11-15, 2022; the *12th International Symposium on Foundations of Information and Knowledge Systems (FoIKS 2022)*, Helsinki, Finland, June 20-23, 2022; the *19th International Conference (ESWC 2022)*, Hersonissos, Crete, Greece, May 29 - June 2, 2022; the *37th Italian Conference on Computational Logic (CILC 2022)*, Bologna, Italy, 29 June - 1 July

2022; the *15th International conference on Scalable Uncertainty Management (SUM 2022)*, Paris, France, 17-19, 2022; the *6th International Joint Conference on Rules and Reasoning (RuleML+RR)*, virtual, 26-28 September 2022.

9. Training Activities

The InfraScience activities include several training sessions and courses on Open Science and Research Data Management topics. These training events and courses addressed various types of audiences, which can be mainly grouped into two categories: research support staff - librarians, repository administrators, legal office staff, ethics committees and so on; and researchers at different career stages, from PhD students to researchers and Principal Investigators.

in particular, the following training activities were organised:

- Research Data Management basics: why it is essential to take care of data, at Center for Instrument Sharing of the University of Pisa (CISUP), 20 January 2022, Webinar;
- Open Science: from theory to practice, XXXVI PhD Cycle: cross-cutting learning activities and other highly qualified activities at the University of Pisa, Academic Year 2021-2022;
- Open Science and Open Access. Why and how to guarantee openness to knowledge, at University of Milano Bicocca, in the context of the SURFICE research project. 23 February 2022, webinar;
- Practical course on FAIR Data Stewardship in Life Science. 15-hour course focused on FAIR data management and stewardship. Dates: 2-4-7-9-11 March 2022. Online
- OpenOrgs - the disambiguation tool for your organization. Improving the quality of data in the scholarly communication ecosystem. Training for OpenAIRE Monitor Institutional Dashboard managers. 28 June 2022, webinar.
- On Open Science and the importance of good Research Data Management, Polytechnic of Milan, MSCA ITN PHAST, 7 September 2022
- Open Science and the importance of data. Why a proper Research Data Management is essential for good science, University of Pisa, Biopham European Joint Degree, 15 September 2022.
- RISIS Tool Demonstration Event – The OpenAIRE Research Graph: an Open Access resource for research on research. 26 October 2022. The training session included a presentation of the OpenAIRE graph and a guided practical session where participants could learn

how to use the OpenAIRE Graph for research and policy-related activities.

- The Data Management Plan in the Horizon Europe template, University of Pisa, 22 November 2022, webinar.

10. Working Groups, Task Forces, & Interest Groups

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

- *Gruppo di Lavoro “Roadmap per la scienza aperta del CNR”* (D. Castelli) – A WG called to develop the roadmap leading to the implementation of open science practices by the National Research Council of Italy.
- *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.
- *GOFAIR Discovery Implementation Network* (A. Bardi) – a GO FAIR consortium called to provide interfaces and other user-facing services for data discovery across disciplines;
- *ISTI IT infrastructure (S2I2S)* (F. Debole) – a WG called to drive the development of the Institute IT and services;
- *ISTI Open Access* (L. Candela) – a WG called to drive the development of the Institute open access and open science policies and practices;

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

- *Helmholtz Metadata Collaboration (HMC)* (D. Castelli) – Crossing the lines between the research fields, the Helmholtz Association decided to strengthen activities in the area of metadata significantly. To do this, it set up the Helmholtz Metadata Collaboration (HMC) with an annual budget of 4.9 million Euros. HMC provides funding for innovative metadata approaches in a competitive selection process. The evaluation is conducted by an international panel of experts.
- *International Scientific Committee of the CCSD* (D. Castelli) – The Center for Direct Scientific Communication (CCSD) is a French organization providing the higher education and research community with the tools needed to archive, disseminate and capitalise on scientific publications and data. The international scientific committee is made of 11 qualified personalities, French and international experts in the fields of open

science, publication, open archives and research data that provide advices and recommendations, contribute to the scientific and technological watch and to the international visibility of the programs of the CCDS, and advise on partnership prospects with third parties.

- *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 Innovation Forum Working Group on Data* (D. Castelli) – The EIC Forum serves as a platform where Member States, Associated Countries, and other stakeholders can exchange experiences and ideas regarding these standards and common data policies. Participants discuss their needs, pinpoint gaps, and explore strategies to address these challenges. During discussions, members identified key obstacles, such as discrepancies in data definitions and data accessibility. The Data working group aims to contribute to overcoming the gap characterising the management of innovation-related data by aligning with the objectives outlined in the New European Innovation Agenda, which advocates for the creation and utilization of comprehensive, comparable datasets and the establishment of a common data repository.
- *EOSC Task Force on Technical Interoperability of Data and Services* (P. Manghi) – a TF taking the EOSC Interoperability Framework (EIF) recommendations around technical architecture as their starting point to help develop the EOSC Core and Exchange as described in the SRIA.
- *EOSC Task Force on Infrastructures for Quality Research Software* (L. Candela) – a TF fostering the development and deployment of tools and services that allow researchers to properly archive, reference, describe with proper metadata, share and reuse research software, as well as to improve their quality, both from the technical and organizational point of view.
- *Gruppo di Lavoro "Roadmap per la scienza aperta del CNR"* (L. Candela) – A WG called to develop the roadmap leading to the implementation of open science practices by the National Research Council of Italy.

- *OpenAIRE AMKE Services and Technologies Standing Committee* (C. Atzori, A. Bardi, M. Baglioni) – a committee providing the strategic framework to define, assess, expand, maintain and improve the OpenAIRE services and enhance their interoperability with international, national, regional, and sub-regional services.

11. Collaborations

Besides the collaborations taking place in the context of the projects, InfraScience established the following collaboration agreements.

Azienda Ospedaliero-Universitaria Pisana (AOUP): (*Jul. 2022-ongoing*) InfraScience formally agreed with AOUP (the principal healthcare institute in the Pisa area) to jointly work in the field of Artificial Intelligence for the automatic identification of peculiar dialogue phases in “training through simulation in neonatology”, and for the early detection of pathologies based on infant cry signal processing. In particular, the activities focus on (a) the design and development of AI techniques, (b) the measurement of the performance of the implemented systems on simulated and real scenarios, (c) the evaluation of the usefulness of the developed products for the AOUP training activities, (d) the integration of heterogeneous devices (from haptic to neural) with the developed software.

Istituto di Geoscienze e Georisorse del Consiglio Nazionale delle Ricerche (IGG-CNR): (Nov. 2022-ongoing) In the context of the ITINERIS Italian PNRR project (project code nr. IR0000032, CUP B53C22002150006), InfraScience formally agreed with IGG-CNR for the scientific supervision of human resources working on Virtual Research Environment Management within the project's Work Package 8 - “Virtual Research Environments and Cross-disciplinary Activities”. The focus of the activity within VREs is the identification, analysis, and interpretation of data and trends in the context of the empirical modelling of complex environmental and ecological systems and the interactions between the geo-sphere and the bio-sphere.

Visual Persistence: (*Mar. 2020-Mar. 2023*) Visual Persistence is a company having its place of business in Truro, United Kingdom, Swanpool Street 3, TR11 3HU, tax code UTR9072632908, represented by its Director Dr Matthew Walsh, focusing on advanced underwater devices for sea monitoring. InfraScience and Visual Persistence formally agreed to develop the Underwater Detector of Moving Object's Sizes software (UDMOS), a computer vision process to detect underwater moving objects (possibly fishes) having size larger than a certain threshold, which was able to trigger a camera recording event. The software was coupled with an underwater device developed and owned by Visual Persistence that maximised the performance of UDMOS through the proper setup of cameras, filters, lenses, lights, and protection structures. InfraScience helped Visual Persistence develop a proof-

of-concept for the UDMOS algorithm and hardware and assessed performance on recorded and live underwater scenarios. This collaboration also produced scientific publications.

12. 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 2022.

During 2022 InfraScience members contributed to the publishing of 44 papers, to the research and development activities of 21 research projects (17 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), EcoScope project (grant agreement No. 101000302), EOSC Future project (grant agreement No. 101017536), EOSC-Pillar project (grant agreement No. 857650), I-GENE project (grant agreement No. 862714), MOVING project (grant agreement No. 862739), OpenAIRE Nexus project (grant agreement No. 101017452), PerformFISH project (grant agreement No. 727610), RISIS 2 project (grant agreement No. 824091), Skills4EOSC project (grant agreement No. 101058527), SoBigData-PlusPlus (grant agreement No. 871042), TAILOR (grant agreement No. 952215).

InfraScience received funding from the European Union's Horizon Europe research and innovation programme under: CODECS project (grant agreement No. 101060179), FAIR-CORE4EOSC project (grant agreement No. 101057264), SoBigData RI PPP project (grant agreement No. 101079043).

References

- [1] M. Arezoumandan, L. Candela, D. Castelli, A. Ghanadrad, D. Mangione, and P. Pagano. Virtual Research Environments Ethnography: a Preliminary Study. In *Proceedings of the 14th International Workshop on Science Gateways, Trento, Italy*, 2022. doi: 10.5281/zenodo.7883104.
- [2] O. Arieli, G. Casini, and L. Giordano, editors. *International Workshop on Non-Monotonic Reasoning 2022*, volume 3197 of *CEUR Workshop Proceedings*, 2022. CEUR-WS.org.
- [3] M. Artini, L. Candela, P. Manghi, and S. Giannini. Repogate: 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. ISBN 978-3-030-39905-4. doi: 10.1007/978-3-030-39905-4_15.
- [4] M. Artini, L. Candela, A. Dell'Amico, A. Molino, S. Giannini, and T. Piccioli. ISTI Open Portal activity report 2022. Technical Report 036, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", 2022.
- [5] M. Artini, S. La Bruzzo, M. De Bonis, and G. Pavone. Openorgs: a tool for the disambiguation of organizations. Technical Report 034, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", 2022.
- [6] M. Assante, L. Candela, D. Castelli, R. Cirillo, 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. doi: 10.1016/j.future.2018.10.035.
- [7] 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, Dec. 2019. ISSN 0167739X. doi: 10.1016/j.future.2019.05.063.
- [8] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, A. Dell'Amico, L. Frosini, L. Lelii, M. Lettere, F. Mangiacrapa, P. Pagano, G. Panichi, T. Piccioli, and F. Sinibaldi. Virtual research environments co-creation: The D4Science experience. *Concurrency and Computation: Practice and Experience*, Mar. 2022. ISSN 1532-0626, 1532-0634. doi: 10.1002/cpe.6925.
- [9] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, A. Dell'Amico, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, T. Piccioli, F. Sinibaldi, and F. Zoppi. D4science activity report 2022. Technical Report 037, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", 2022.
- [10] M. Baglioni, A. Bardi, C. Atzori, and P. Manghi. Books from the OpenAIRE Research Graph, June 2022.
- [11] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, Apr. 2022.
- [12] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, May 2022.
- [13] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, June 2022.

- [14] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, June 2022.
- [15] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, July 2022.
- [16] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, Sept. 2022.
- [17] M. Baglioni, A. Bardi, H. Dimitropoulos, C. Atzori, and P. Manghi. OpenAIRE Research Graph Dump: new collected projects, Nov. 2022.
- [18] A. Bardi, I. Kuchma, G. Pavone, M. Artini, C. Atzori, A. Bäcker, M. Baglioni, A. Czerniak, M. De Bonis, H. Dimitropoulos, I. Fofoulas, M. Horst, K. Iatropoulou, P. Jacewicz, A. Kokogiannaki, S. La Bruzzo, E. Lazzeri, A. Löhden, P. Manghi, A. Mannocci, N. Manola, E. Ottonello, and J. Schirrwagen. OpenAIRE Covid-19 publications, datasets, software and projects metadata., Dec. 2022.
- [19] A. Berti, G. Carloni, S. Colantonio, M. A. Pascali, P. Manghi, P. Pagano, R. Buongiorno, E. Pachetti, C. Caudai, D. Di Gangi, E. Carlini, Z. Falaschi, E. Ciarracchi, E. Neri, E. Bertelli, V. Miele, R. Carpi, G. Bagnacci, N. Di Meglio, M. A. Mazzei, and A. Barucci. Data Models for an Imaging Bio-bank for Colorectal, Prostate and Gastric Cancer: the NAVIGATOR Project. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04, Ioannina, Greece, Sept. 2022. IEEE. ISBN 978-1-66548-791-7. doi: 10.1109/BHI56158.2022.9926910.
- [20] R. Borgheresi, A. Barucci, S. Colantonio, G. Aghakhanyan, M. Assante, E. Bertelli, E. Carlini, R. Carpi, C. Caudai, D. Cavallero, D. Cioni, R. Cirillo, V. Colcelli, A. Dell’Amico, D. Di Gangi, P. A. Erba, L. Faggioni, Z. Falaschi, M. Gabelloni, R. Gini, L. Lelii, P. Liò, A. Lorito, S. Lucarini, P. Manghi, F. Mangiacrapa, C. Marzi, M. A. Mazzei, L. Mercatelli, A. Mirabile, F. Mungai, V. Miele, M. Olmastroni, P. Pagano, F. Paiar, G. Panichi, M. A. Pascali, F. Pasquinelli, J. E. Shortrede, L. Tumminello, L. Volterrani, E. Neri, and on behalf of the NAVIGATOR Consortium Group. NAVIGATOR: an Italian regional imaging biobank to promote precision medicine for oncologic patients. *European Radiology Experimental*, 6(1):53, Nov. 2022. ISSN 2509-9280. doi: 10.1186/s41747-022-00306-9.
- [21] A. Burton, A. Aryani, H. Koers, P. Manghi, S. La Bruzzo, M. Stocker, M. Diepenbroek, U. Schindler, and M. Fenner. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine*, 23 (1/2), Jan. 2017. ISSN 1082-9873. doi: 10.1045/january2017-burton.
- [22] L. Candela, D. Castelli, and D. Mangione. Comparison of federated solutions for distributed infrastructures. Technical Report 024, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [23] L. Candela, D. Castelli, and D. Mangione. Research infrastructures: an open science quandary. Technical Report 025, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [24] L. Candela, D. Castelli, and D. Mangione. Research workflows and open science. Technical Report 026, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [25] G. Casini and U. Straccia. Defeasible reasoning in RDFS. In O. Arieli, G. Casini, and L. Giordano, editors, *International Workshop on Non-Monotonic Reasoning 2022*, volume 3197 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
- [26] G. Casini and U. Straccia. A General Framework for Modelling Conditional Reasoning - Preliminary Report. In *Proceedings of the Nineteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 112–121, Haifa, Israel, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-01-0. doi: 10.24963/kr.2022/12. URL <https://proceedings.kr.org/2022/12>.
- [27] G. Casini and U. Straccia. A Rational Entailment for Expressive Description Logics via Description Logic Programs. In E. Jembere, A. J. Gerber, S. Viriri, and A. Pillay, editors, *Artificial Intelligence Research*, volume 1551, pages 177–191. Springer International Publishing, Cham, 2022. ISBN 978-3-030-95069-9 978-3-030-95070-5. doi: 10.1007/978-3-030-95070-5_12. Series Title: Communications in Computer and Information Science.
- [28] G. Casini, T. Meyer, G. Paterson-Jones, and I. Varzinczak. Klm-style defeasibility for restricted first-order logic. In G. Governatori and A.-Y. Turhan, editors, *Rules and Reasoning*, pages 81–94, Cham, 2022. Springer International Publishing. ISBN 978-3-031-21541-4.
- [29] G. Casini, T. Meyer, and I. Varzinczak. Situated conditionals - a brief introduction. In O. Arieli, G. Casini, and L. Giordano, editors, *International Workshop on Non-Monotonic Reasoning 2022*, volume 3197 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
- [30] G. Casini, L. Robaldo, L. van der Torre, and S. Villata, editors. *Handbook of Legal AI*, 2022. College Publications.
- [31] G. Coro and P. Bove. A High-resolution Global-scale Model for COVID-19 Infection Rate. *ACM Transactions*

- on *Spatial Algorithms and Systems*, 8(3):1–24, Sept. 2022. ISSN 2374-0353, 2374-0361. doi: 10.1145/3494531.
- [32] G. Coro, S. Bardelli, A. Cuttano, and N. Fossati. Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology. *Education and Information Technologies*, 27(7): 9181–9203, Aug. 2022. ISSN 1360-2357, 1573-7608. doi: 10.1007/s10639-022-11000-z.
- [33] G. Coro, P. Bove, E. N. Armelloni, F. Masnadi, M. Scanu, and G. Scarcella. Filling Gaps in Trawl Surveys at Sea through Spatiotemporal and Environmental Modelling. *Frontiers in Marine Science*, 9:919339, July 2022. ISSN 2296-7745. doi: 10.3389/fmars.2022.919339.
- [34] G. Coro, P. Bove, and A. Ellenbroek. Habitat distribution change of commercial species in the Adriatic Sea during the COVID-19 pandemic. *Ecological Informatics*, 69: 101675, July 2022. ISSN 15749541. doi: 10.1016/j.ecoinf.2022.101675.
- [35] G. Coro, A. N. Tassetti, E. N. Armelloni, J. Pulcinella, C. Ferrà, M. Sprovieri, F. Trincardi, and G. Scarcella. COVID-19 lockdowns reveal the resilience of Adriatic Sea fisheries to forced fishing effort reduction. *Scientific Reports*, 12(1):1052, Jan. 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-05142-w.
- [36] M. De Bonis, P. Manghi, and C. Atzori. FDup: a framework for general-purpose and efficient entity deduplication of record collections. *PeerJ Computer Science*, 8: e1058, Sept. 2022. ISSN 2376-5992. doi: 10.7717/peerjcs.1058.
- [37] D. Farace, S. Biagioni, C. Carlesi, and C. Baars. Persistent identifiers and grey literature: A PID project and greynet use case. In M. Leonard, S. E. Thomas, and Core, editors, *Managing Grey Literature: Technical Services Perspectives*. ALA Editions Core, 2022.
- [38] P. O. Garcia, L. Berberi, L. Candela, I. Van Nieuwerburgh, E. Lazzeri, and M. Czurray. Developing the EOSC-Pillar RDM Training and Support Catalogue. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, and A. Poggi, editors, *Linking Theory and Practice of Digital Libraries*, volume 13541, pages 274–281. Springer International Publishing, Cham, 2022. ISBN 978-3-031-16801-7 978-3-031-16802-4. doi: 10.1007/978-3-031-16802-4_22. Series Title: Lecture Notes in Computer Science.
- [39] A. Ghannadrad, M. Arezoumandan, L. Candela, and D. Castelli. Recommender systems for science: A basic taxonomy. In G. M. Di Nunzio, B. Portelli, D. Redavid, and G. Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
- [40] S. La Bruzzo and P. Manghi. OpenAIRE ScholXplorer Service: Scholix JSON Dump, Mar. 2022.
- [41] S. La Bruzzo and P. Manghi. Scholexplorer activity report 2022. Technical Report 035, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [42] S. La Bruzzo, M. Artini, C. Atzori, A. Bardi, M. Baglioni, M. De Bonis, A. Mannocci, P. Manghi, and G. Pavone. Data model description of the openaire research graph. Technical Report 031, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [43] S. La Bruzzo, M. Artini, C. Atzori, A. Bardi, M. Baglioni, M. De Bonis, A. Mannocci, P. Manghi, and G. Pavone. Openaire research graph deduplication workflow. Technical Report 032, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [44] S. La Bruzzo, M. Artini, C. Atzori, A. Bardi, M. Baglioni, M. De Bonis, A. Mannocci, P. Manghi, and G. Pavone. Openaire research graph: aggregation workflow. Technical Report 033, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [45] 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. doi: 10.1108/PROG-08-2013-0045.
- [46] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, and T. Vergoulis. New trends in scientific knowledge graphs and research impact assessment. *Quantitative Science Studies*, 2(4):1296–1300, Dec. 2021. ISSN 2641-3337. doi: 10.1162/qss.e_00160.
- [47] P. Manghi, M. Artini, S. La Bruzzo, E. Ottonello, and G. Pavone. Open science repository platforms. Technical Report 009, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [48] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Czerniak, M. Horst, K. Kiatropoulou, A. Kokogianaki, M. De Bonis, M. Artini, A. Lempesis, A. Mannocci, A. Ioannidis, T. Vergoulis, S. Chatzopoulos, and D. Pierrakos. OpenAIRE Research Graph: Dump of funded products, Dec. 2022.
- [49] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Czerniak, M. Horst, K. Kiatropoulou, A. Kokogianaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis,

- A. Mannocci, and A. Ioannidis. OpenAIRE Research Graph: Dump of funded products, Mar. 2022.
- [50] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Czerniak, M. Horst, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Mannocci, and A. Ioannidis. OpenAIRE Research Graph: Dump of funded products, June 2022.
- [51] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Löhden, A. Bäcker, A. Mannocci, M. Horst, A. Czerniak, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, and F. Summan. OpenAIRE Research Graph: Dumps for research communities and initiatives., Mar. 2022.
- [52] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Löhden, A. Bäcker, A. Mannocci, M. Horst, A. Czerniak, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, and F. Summan. OpenAIRE Research Graph: Dumps for research communities and initiatives., June 2022.
- [53] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Löhden, A. Bäcker, A. Mannocci, M. Horst, A. Czerniak, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, and F. Summan. OpenAIRE Research Graph: Dumps for research communities and initiatives., Dec. 2022.
- [54] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Mannocci, M. Horst, A. Czerniak, K. Iatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, A. Lempesis, A. Ioannidis, N. Manola, P. Principe, T. Vergoulis, S. Chatzopoulos, and D. Pierrakos. OpenAIRE Research Graph Dump, Dec. 2022.
- [55] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Mannocci, M. Horst, A. Czerniak, K. Iatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, N. Manola, and P. Principe. OpenAIRE Research Graph Dump, June 2022.
- [56] D. Mangione, L. Candela, and D. Castelli. A taxonomy of tools and approaches for fairification. In G. M. Di Nunzio, B. Portelli, D. Redavid, and G. Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
- [57] A. Mannocci, M. Baglioni, and P. Manghi. “Knock Knock! Who’s There?” A Study on Scholarly Repositories’ Availability. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, and A. Poggi, editors, *Linking Theory and Practice of Digital Libraries*, volume 13541, pages 306–312. Springer International Publishing, Cham, 2022. ISBN 978-3-031-16801-7 978-3-031-16802-4. doi: 10.1007/978-3-031-16802-4_26. Series Title: Lecture Notes in Computer Science.
- [58] A. Mannocci, O. Irrera, and P. Manghi. Will open science change authorship for good? towards a quantitative analysis. In G. M. Di Nunzio, B. Portelli, D. Redavid, and G. Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
- [59] A. Mannocci, O. Irrera, and P. Manghi. Open science and authorship of supplementary material. evidence from a research community. In N. Robinson-Garcia, D. Torres-Salinas, and W. Arroyo-Machado, editors, *26th International Conference on Science, Technology and Innovation Indicators (STI 2022)*. 2022. doi: <https://doi.org/10.5281/zenodo.6975411>.
- [60] J. Maranhão, G. Casini, G. Pigozzi, and L. van der Torre. Normative change: An AGM approach. *Journal of Applied Logics - IfCoLog Journal*, 9(4), 2022.
- [61] F. Minutella, F. Falchi, P. Manghi, M. De Bonis, and N. Messina. Towards unsupervised machine learning approaches for knowledge graphs. In G. M. Di Nunzio, B. Portelli, D. Redavid, and G. Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
- [62] E. Ottonello, M. Artini, S. La Bruzzo, and G. Pavone. Bioschemas data sources aggregation to openaire research graph. Technical Report 010, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, 2022.
- [63] G. Scarcella, S. Angelini, E. N. Armelloni, I. Costantini, A. De Felice, S. Guicciardi, I. Leonori, F. Masnadi, M. Scanu, and G. Coro. The potential effects of COVID-19 lockdown and the following restrictions on the status of eight target stocks in the Adriatic Sea. *Frontiers in Marine Science*, 9:920974, Oct. 2022. ISSN 2296-7745. doi: 10.3389/fmars.2022.920974.
- [64] D. Schaap, M. Assante, P. Pagano, and L. Candela. Blue-Cloud: Exploring and demonstrating the potential of Open Science for ocean sustainability. In *2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*, pages 198–202, Milazzo, Italy, Oct. 2022. IEEE. ISBN

978-1-66549-942-2. doi: 10.1109/MetroSea55331.2022.9950819.

- [65] U. Straccia and G. Casini. A Minimal Deductive System for RDFS with Negative Statements. In *Proceedings of the Nineteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 351–361, Haifa, Israel, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-01-0. doi: 10.24963/kr.2022/35. URL <https://proceedings.kr.org/2022/35>.
- [66] C. Thanos, C. Meghini, V. Bartalesi, and G. Coro. An exploratory approach to archaeological knowledge production. *International Journal on Digital Libraries*, 23 (3):231–239, Sept. 2022. ISSN 1432-5012, 1432-1300. doi: 10.1007/s00799-022-00324-3.
- [67] T. Vergoulis, S. Chatzopoulos, K. Vichos, I. Kanellos, A. Mannocci, N. Manola, and P. Manghi. BIP! SCHOLAR: A Service to Facilitate Fair Researcher Assessment. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5, Cologne Germany, June 2022. ACM. ISBN 978-1-4503-9345-4. doi: 10.1145/3529372.3533296.
- [68] K. Vichos, M. De Bonis, I. Kanellos, S. Chatzopoulos, C. Atzori, N. Manola, P. Manghi, and T. Vergoulis. A preliminary assessment of the article deduplication algorithm used for the openaire research graph. In G. M. Di Nunzio, B. Portelli, D. Redavid, and G. Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

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 Senior Technology Researcher (Primo Tecnologo) at the “Istituto di Scienza e Tecnologie della Informazione A. Faedo” (ISTI), an institute of the Italian National Research Council (CNR). He has strong theoretical foundation in Computer Science and Technology, encompassing mathematics, logic, and cross-platform coding, and over 15 years of experience working on distributed systems, e-infrastructures and Virtual Research Environments. Massimiliano’s academic journey includes a Ph.D. in Information Engineering, a Master’s degree (M.Sc.) in Information

Technologies, and a Bachelor’s degree (B.Sc.) in Computer Science, all earned from the University of Pisa. Throughout his career, Massimiliano has been actively involved in several European Union (EU) projects, including ARIADNE-Plus, Blue-Cloud, AGINFRAPPlus, BlueBRIDGE, DESIRA, PARTHENOS, SoBigData, EOSCPilot, PerformFISH, iMarine, EU-BrazilOpenBio, D4Science II, D4Science, and DILIGENT. He has held various roles in these projects, showcasing his versatility and adaptability. Currently, Massimiliano is engaged in multiple EU projects with responsibility roles, including Blue-Cloud 2026, FOSSR PNRR, SoBigData-PlusPlus, MOVING, and RISIS2.


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 product manager role 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.


Pasquale Bove is a PhD researcher in computer science at the Institute of Information Science and Technologies of the Italian National Research Council. His research focuses on Data Mining and Ecological Niche Modeling. His work is currently focused on the experimentation of models and


methodologies to process biological and environmental data, especially in the marine field, with an Open Science and science reproducibility-oriented approach.

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 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 team. Under her supervision, the InfraScience team coordinated and participated in several EU and nationally funded projects on Digital Libraries and Research Data Infrastructures. She has participated as an expert to the shaping of the Italian National Plan for Open Science and 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 authored 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 BlueCloud EU Project. He was involved in various EU-funded projects including BlueBridge, 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 ma-

chine 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.

Michele De Bonis  is a research fellow at the Institute of Science and Information Technologies 'A Faedo' (ISTI) of the CNR of Pisa, and a PhD student of Information Engineering at the University of Pisa. He graduated in Computer Science at the University of Pisa and his research focuses on entity deduplication on big scholarly communication graphs. In particular, the aim of his studies is to find solutions for author name disambiguation and entity linking based on Artificial Intelligence and Deep Learning techniques. Michele joined ISTI in 2017 and he worked for the projects in the OpenAIRE Infrastructure.

Franca Debole  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 a technical director and involved in several European and National projects. Her current research activities range from digital image processing to techniques for image retrieval and automatic annotation tools. 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. 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 Schoexpler 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).

Lucio Lelii is a 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.

Paolo Manghi is a (PhD) Researcher in computer science at Istituto di Scienza e Tecnologie dell’Informazione (ISTI) of Consiglio Nazionale delle Ricerche (CNR), in Pisa, Italy. He is the CTO of OpenAIRE AMKE, involved in coordination and/or activities in the H2020 projects FAIRCORE4EOSC, EOSC-Future, EOSC-Enhance, OpenAIRE-Nexus, OpenAIRE-Connect, OpenAIRE-Advance, OpenAIRE2020. His research areas of interest are today 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.

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 a background in 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 involves the design and development of (Web-)GUI based on several frameworks (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 (BlueCloud, SoBigData, PARTHENOS) and is responsible for: Data Access and Exchange (Workspace Area), Data Catalogue and Publishing (Catalogue Area).

Dario Mangione is a library and information scientist and a graduate fellow at the National Research Council of Italy, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”. His research activity is focused on the study and development of models, solutions, and systems enabling and fostering open and Findable, Accessible, Interoperable, and Reusable (FAIR) practices and ultimately an open science approach. His research interests include semantic Web oriented controlled vocabularies and metadata standards. He has been involved as a terminology expert in the EC-funded EOSC-Secretariat.eu project for supporting the development of standardisation solutions within the scope of the creation of the European Open Science Cloud (EOSC). He is currently work-

ing on FAIR digital objects evaluation practices.

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.

Enrico Ottonello is a research fellow at Istituto di Scienza e Tecnologie dell’Informazione A. Faedo (ISTI), an institute of the Italian National Research Council (CNR), since December 2018. He graduated in computer science at the University of Pisa in 2002. He played the role of software engineer for several companies from 2003 up to 2018. At ISTI-CNR he worked for several EU Projects including OpenAIRE-Advance and OpenAIRE-Connect. His research interests and activities focus on data cleaning, anomaly detection, and enrichment in scholarly communication graphs.

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. 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 including iMarine, BlueBRIDGE, EUBrazilOpenBio and ENVRI. He is currently mainly involved in BlueCloud and EOSC-Pillar projects.

Gina Pavone is a research fellow focusing on Open Science, Open Access and Research Data Management. She is in charge as National Open Access Desk of OpenAIRE and she coordinates the editorial board of open-science.it, the Ital-

ian portal dedicated to the many components of Open Science. She is also member OpenAIRE Community of Practice of Training Coordinators and she is involved in the structuring of a national Competence Centre for Open Science, FAIR data and EOSC within the ICDI (Italian Computing and Data Infrastructure). Her activities range from the definition of strategies and tools for the support and training of researchers to the dissemination of Open Science activities and initiatives. She has worked in several international projects such as OpenAIRE, EOSC Pillar, EOSC Secretariat and RDA Europe. She is a journalist with expertise in data analysis, she holds a master's degree in publishing and journalism at the Sapienza University of Rome and a second-level master's degree in big data analytics and social mining at the University of Pisa. She has been involved in campaigns for open data and transparency in public institutes and administrations and she has worked as a data analyst and data journalist for the European Data Journalism Network (EDJNet).

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, SoBig-DataPlus, 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 designing and developing distributed environments' services aimed at managing scientific data, with special attention to Ecological Niche Modelling approaches. These studies involved

the 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. In the past he has been involved in the iMarine, EAGLE, EUBrazilOpenBio, ENVRI, Venus-C, D4Science-II, D4Science projects.

Umberto Straccia is 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.

Franco Zoppi is Associated Researcher at CNR-ISTI. He graduated in Computer Science at the University of Pisa and has been working since the '80s on the design and implementation of software systems in the areas of DBMS, Distributed Office Information Systems and Digital Library Systems. Initially employed at the Research and Development Department of Olivetti S.p.A., then at the Network Laboratory of the Telecommunications Department of Telecom Italia, in 2001 he joined the Information System Department of Pisa University as Project Manager. Since 2005 he has been working as a Researcher at CNR-ISTI, where he coordinated the CNR activities in the BELIEF/BELIEF-II, HOPE and EAGLE projects and was involved in the management of several EC projects such as D4Science, BlueBRIDGE, RDA Europe, DRIVER I/II, OpenAIRE, OpenAIREplus, OpenAIRE2020, EFG. He is currently involved in the management of most of the InfraScience group projects.