14th International Workshop on Science Gateways (IWSG 2022), 15-17 June 2022

Virtual Research Environments Ethnography: a Preliminary Study

M. Arezoumandan, L. Candela, D. Castelli, A. Ghannadrad, D. Mangione, P. Pagano
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
National Research Council of Italy,
via G. Moruzzi, 1, Pisa, Italy
Email: {name.surname}@isti.cnr.it

***Abstract*—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.**

***Keywords*—Virtual Research Environment; Science Gateway; Virtual Laboratory; Survey and overview; Systematic literature review**

I. Introduction

Science gateways (SGs) [1], *Virtual Research Environments* (VREs) [2] and *Virtual Laboratories* (VLabs) are all terms used to indicate solutions aiming at providing a designated community with online research platform catering for integrated access to *resources* (e.g. computing, software, data, instruments) of interest for the community [3], [4]. However, the scope of research studies under this definition is ample and varied.

This paper presents the first systematic mapping study on literature about this family of systems and solutions. We retrieve and select 1167 research papers from the literature in the period 2010-2022 to systematically analyse this corpus and identify significant trends and characteristics. In particular, the study aims at identifying whether there are intrinsic differences among studies classified as SGs with respect to VREs or VLabs and vice versa. The study focus on three aspects: the distribution of studies over time, the publication venues and the characterising topics.

The remainder of the paper is organised as follows. Sec. II describes related reviews to motivate this study. Sec. III reports the method exploited by the study. Sec. IV presents the early results and discusses the threats to validity. Finally, Sec. V concludes the paper and discusses future works.

II. Related works

Several studies have been published aiming to describe the state of the art of SGs, VLabs, and/or VREs.

Lawrence et al. [1] conducted an extensive survey with 5000 respondents including principal investigators, senior administrators, and people with gateway affiliations. That survey indicated that SGs were an active part of the science and engineering research and education landscape. According to the study, among the various tools and services SGs offer, education tools, computational tools, data analysis tools, and data collections were the most common. The resources SGs give access to are provided in different ways including in house development, the provisioning of computational tools and data collections by public or academic institutions, the acquisition of collaboration tools, scientific instruments, and rapid publishing mechanisms by commercial providers. Moreover, the study highlighted that SGs served all sizes of communities, and this has implications for technology, staffing, and development methods. For instance, SGs are called to serve simple cases where tens to thousands of students are provided with online, course-integrated environments to support data analysis and computational experiments for the duration of a course up to discipline-specific gateways called to serve entire science communities, thus requesting developers, operators, and support personnel for the long term.

Barker et al. [3] discussed some definitions of the three terms Science gateways, Virtual Research Environments and Virtual Laboratories existing in the literature. In particular, the study highlighted the origins of the different terms and reported some features characterising the three, concluding that these terms were actually referring to similar classes of systems. It also clarified that these systems differentiate from generic cyberinfrastructures or digital (research) infrastructures on which they can be built. The value these systems should bring regards "lowering barriers to infrastructures, enabling collaboration between (remote) researchers and across multiple disciplines, sharing and linking infrastructure resources, driving standards and open science, and supporting teaching and new career developments.".

Calyam et al. [4] analysed an array of VREs and SGs initiatives by using the Science Gateways Community Institute clientele and the International Virtual Research Environment Interest Group of the Research Data Alliance. The goal of the study was to collect metrics and indicators suitable for characterising the impact of these initiatives. A rich set of approaches to routinely measure and communicate impact

Corresponding author leonardo.candela@isti.cnr.it were identified, leading to the identification of four primary areas: user type and count, user behaviour, user satisfaction, and long-term impacts.

Sepúlveda-Rodríguez et al. [5] identified and analysed 168 primary studies on frameworks, models, methodologies, processes, and good practices to manage IT resources and services to realise Science Gateways. This study concludes by recommending the exploitation of cloud technologies to guarantee an "adequate management of the set of IT resources and services used to support Science Gateway environments".

All these studies tend to agree on the convergence of the systems originating from the three classes into a common one.

Diwakar et al. [6] and Panasiuk et al. [7] discussed the role virtual laboratories have in engineering education. They discussed advantages (e.g. cheaper than real laboratories, convenient for dangerous experiments) and disadvantages (e.g. impossibility to completely replace a real experiment with a computer one) of such virtual laboratories concluding that they are an effective tool for practical learning.

Environments for supporting education share features with the rest of SGs, VREs and VLabs, yet they have their own peculiarities.

Our investigation aim at identifying commonalities and differences among the studies on SGs, VREs and VLabs by systematically collecting and analysing the literature published in the last 12 years.

III. Methodology

This research was carried out as a Systematic Mapping Study (SMS) [8], [9] to answer three research questions:

RQ1: what is the distribution over time of the literature on virtual research environments?

RQ2: what are the most relevant journals and conferences on the subject of virtual research environments?

RQ3: which topics can be identified within the scope of virtual research environments, and what is their distribution?

We selected ACM, IEEEXplore, ScienceDirect, Scopus, and Springer databases for conducting the literature search, and we identified five relevant keywords, distributed into three groups, to be used for formulating queries, reflecting the different terms used for designating our subject of interest in different contexts: (*i*) "virtual research environment"; (*ii*) "virtual laboratory" OR "vlab"; (*iii*) "science gateway" OR "scientific gateway". Following a preliminary analysis, the abbreviations VRE and SG were discarded to reduce the noise. We formulated the queries accordingly, taking into consideration the possible distinction between plural and singular forms. The selected terms were used to develop search strings for retrieving papers matching them on title, abstract, or keywords across the selected databases.

By splitting the queries according to the three groups of keywords, we managed to identify 7775 entries, further limited to 1167. Being interested in the consolidated literature, we removed the entries without a DOI and the duplicates. We also restricted our corpus to journal publications and conference proceedings published since 2010, following a significant peak in the yearly entry distribution: we observed a mean of 51.5 papers per year before 2010, while the mean rose to 217.4 from 2010 onward. Finally, we performed a manual analysis, by reading the abstracts and selected parts of the papers, to further assess the relevance of the entries to our research objectives.

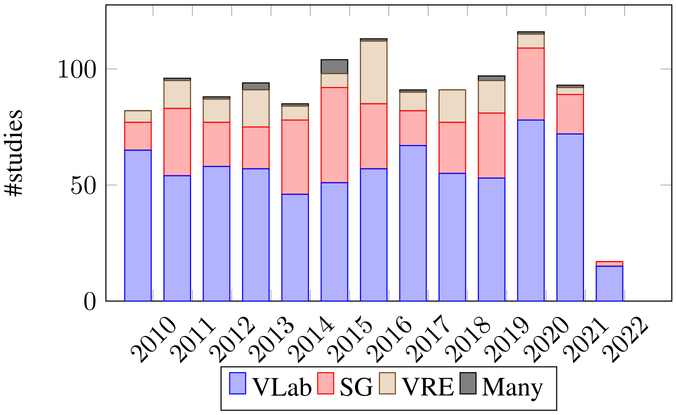

Fig. 1. Studies by year.

IV. Results and Discussion

In the following, we report the results of our analysis.

A. Studies by year

Fig. 1 depicts the distribution of studies by year. It highlights that 62% circa of studies in the sample are on VLabs, 25% circa are on SGs, and 10% circa are on VREs. Only 1% circa of the studies (18 up to 1167) have been characterised by two or more classes with just two studies (i.e. [3] and [10]) using all of them.

Most of the virtual laboratory papers were published in recent years, 78 in 2020 and 72 in 2021. In the case of science gateway, out of 294 studies, 41 papers were published in 2015, which is the highest number of papers among the other years. The highest number of papers regarding virtual research environments, 27 out of 127, was published in 2016, of which 21 are conference papers. It is also worth mentioning that 6 out of 18 papers that have been annotated with a combination of keywords pertaining to more than one of the previously defined groups (SG, VRE, and VLab) were published in 2015, which is the highest rate observed among the other years.

B. Venue

To answer RQ2, we defined the relevance threshold first, opting for the three-sigma limit, a statistical method for filtering 99.7% (or $\geq$ 88.8% in the case of a non-normal distribution) of the values lying in a range of $\mu \pm 3\sigma$, where μ is the mean of the values and σ is the standard deviation, which measures the dispersion of those values. The rationale behind this choice is that we needed a non-arbitrary and restrictive threshold, that could help us identify the most significant values among the ones representing journal and conference venues. Consequently, we divided our dataset between journal

TABLE I
MOST FREQUENT JOURNALS BY STUDIES

Journal name	SG	VLab	VRE	Total
Concurrency Computat Pract Exper	**42**	1	**8**	**51**
Comput Appl Eng Educ	0	**35**	0	**35**
Future Generation Computer Systems	**19**	4	**3**	**26**
iJOE	0	**24**	0	**24**
Journal of Grid Computing	**19**	0	1	**20**
Biodiversity Data Journal	0	1	**3**	4
Data Science Journal	1	0	**3**	4

TABLE II
MOST FREQUENT CONFERENCES BY STUDIES

Conference name	SG	VLab	VRE	Total
PEARC	**55**	0	0	**55**
EDUCON	0	**23**	0	**23**
XSEDE	**20**	0	0	**20**
TeraGrid Conference	**15**	1	0	**16**
IFAC ACE	0	**16**	0	**16**
International Conference on e-Science	8	1	**7**	**16**
Gateway Computing Env. Work.	**14**	0	0	**14**
Int. Work. on Science Gateways	**14**	0	0	**14**
REV	0	**11**	0	11
Frontiers in Education	0	**9**	0	9
IEEE ICETA	0	**8**	0	8
IEEE T4E	0	**8**	0	8
ITHET	0	**7**	0	7
ICL	0	**7**	0	7
IFAC World Congress	0	**7**	0	7
AVI Work. on Big Data Applications	0	0	**6**	6
Int. Conf. on Computational Science	2	1	**3**	6
ACM/IEEE JCDL	0	0	**3**	3
3D Res. Chal. in Cultural Heritage	0	0	**3**	3
Archiving Conference	0	0	**3**	3

and conference papers, and we applied the three-sigma limit to each of the columns representing the number of papers published in those venues with regard to the three different groups of keywords and their sum, which we used as an indicator of the overall relevance of the venue. We considered relevant only the venues where $n > \mu + 3\sigma$, where n is equal to the number of relevant papers we observed.

The results showed that there are five most relevant journals (see the embolden values in Tab. I, column "Total"), over a total of 207 distinct venues ($n > 17.1$), and that the most relevant venues for each group (the embolden values in the column SG, VLab and VRE, with $n > 11$, 10.6, and 2.7 respectively) tend to correspond to the five most relevant ones but two, which are both top VRE venues. Moreover, while there are common venues among the most relevant ones for SGs and VREs, we can observe that the top VLab venues focus on education (Computer Applications in Engineering Education) and engineering (International Journal of Online and Biomedical Engineering).

With respect to conferences, we could observe the eight most relevant venues (see the embolden values in Tab. II, column "Total") over a total of 382 entries ($n > 13.1$), which encompass the five most relevant SG venues respectively, two out of the nine most relevant VLab venues, and one out of the six most relevant VRE venues. The most relevant venues for each group (the embolden values in the column SG, VLab and VRE, with $n > 10.4$, 6.9, and 2.3, respectively) tend to distinguish the terms used, with the VLab ones using only "virtual laboratory" or "VLab", the SG ones referring only to "science gateway" or "scientific gateway", with the exception of just one reference to "virtual laboratory" or "vlab", and the VRE ones mainly using "virtual research environment". When looking at the distribution of venues per group, we could observe that the venues dedicated to educational aspects of our research topic are limited within the scope of virtual laboratories, where they represent the majority of the entries. While this is less evident in the case of journals, with Computer Applications in Engineering Education being just one out of the two relevant VLab venues, it is a clear trend among the conferences, with the IEEE Global Engineering Education Conference, the IFAC Symposium on Advances in Control Education, the Frontiers in Education, the IEEE International Conference on Emerging eLearning Technologies and Applications, the IEEE International Conference on Technology for Education, the International Conference on Information Technology Based Higher Education and Training, and the International Conference on Interactive Collaborative Learning being eight educational oriented venues out of a total of nine VLab venues.

C. Topics

We used a topic modelling approach [11] to extract topics from our corpus. Topic modelling is an unsupervised machine learning method that is capable of scanning a collection of papers, detecting words within them, and automatically clustering word groups and similar phrases that best represent a set of papers. In this project, we applied latent Dirichlet allocation (LDA) [12], one of the most used topic modelling techniques. First, we split the dataset into three datasets: (*i*) the "Science Gateway" dataset containing the studies retrieved by the corresponding terms, (*ii*) the "Virtual Research Environment" dataset containing studies retrieved by the corresponding terms, and (*iii*) the "Virtual Laboratory" dataset containing studies retrieved by the corresponding terms. After creating the datasets, we performed text pre-processing techniques such as tokenisation and lemmatisation on the abstract, keywords and title of each paper. One of the most challenging parts of the LDA is to choose the number of topics (K). Our chosen approach for finding the optimal number of topics was to create many LDA models with different values of the number of topics (K) and choose the one that offered the highest coherence value. Thus, we trained our LDA models on the abstract, keywords and title of each paper with 2, 8, 14, 20, 26, 32, and 38 topics. We then employed the elbow method to determine the optimum K obtaining: 14 for the SG dataset, 8 for the VRE dataset, and 14 for the VLab dataset.

Tables III, IV, and V report the number of topics for each dataset, the selected keywords distribution within the topics, and the number of papers falling into each topic.

Regarding SG topics (Tab. III), none of the identified is prevalent. Some families of related topics might be identified through data visualisation, namely (*i*) SG3 and SG12 are on

TABLE III
TOPICS CHARACTERISING SCIENCE GATEWAYS SELECTED STUDIES

Topic	# Studies	Top 30 representative terms
SG1	26	system, datum, analysis, big, base, architecture, computation, implement, high, application, discovery, challenge, national, computer, complexity, improve, compute, bioinformatics, requirement, result, biology, efficient, meet, show, sequence, usability, propose, expert, major, management
SG2	12	web, include, execution, code, image, tool, feature, software, application, file, server, framework, compute, effort, module, generate, manage, processing, student, output, remote, integrate, rapid, ability, collection, language, client, local, generation, target
SG3	32	model, datum, environment, framework, community, enable, geospatial, modeling, challenge, analytic, develop, address, impact, cybergis, domain, scalable, climate, capability, integrate, build, hubzero, sustainability, earth, face, leverage, building, galaxy, significant, business, due
SG4	26	datum, storage, data, processing, dataset, large, process, portal, analysis, source, visualization, pipeline, analyze, database, archive, domain, time, make, search, transfer, information, current, amount, product, handle, scale, astronomy, explore, host, collect
SG5	21	support, simulation, molecular, analysis, study, base, result, describe, number, structure, detail, process, interaction, author, experiment, method, parameter, identify, extend, combine, mosgrid, protein, structural, laboratory, basis, form, computational, submit, seagrid, performance
SG6	21	service, application, provide, web, science, base, interface, management, development, compute, software, component, developer, capability, build, design, platform, architecture, specific, introduce, api, deploy, advanced, airavata, browser, authentication, authorization, rich, identity, add
SG7	20	gateway, science, cloud, platform, integration, describe, software, implementation, experience, present, solution, paper, development, middleware, deployment, design, generic, cost, scigap, multi, configuration, instance, simplify, host, involve, reference, requirement, creation, scenario, provider
SG8	14	gateway, scientific, present, paper, open, project, discuss, end, teragrid, scientist, management, problem, manage, future, focus, order, life, standard, production, ultrascan, potential, multiple, program, challenge, advantage, engineering, key, heterogeneous, collaboration, energy
SG9	26	research, science, researcher, support, share, community, collaboration, group, repository, metadata, network, access, social, collaborative, set, knowledge, technology, digital, exist, team, discipline, practice, make, enhance, information, deploy, increase, institution, offer, level
SG10	26	infrastructure, science, grid, user, distribute, compute, environment, portal, technology, gateway, virtual, interface, security, cluster, middleware, access, desktop, common, concept, usage, aim, develop, goal, operation, case, mechanism, community, international, build, single
SG11	17	computing, computational, hpc, tool, high performance, work, gateway, require, project, design, neuroscience, science, parallel, complex, barrier, demand, high, develop, successful, consist, center, grow, software, successfully, make, insight, association, machine learne, large scale, addition
SG12	19	user, simulation, base, run, approach, acm, tool, develop, computational, resource, time, execute, online, serve, learn, utilize, education, job, material, system, variety, bring, submission, result, portal, area, learning, multiple, benefit, ieee
SG13	29	workflow, scientific, science, tool, integrate, framework, enable, task, visualization, application, interactive, 'ws-pgrade', scientist, ieee, paper, distribute, guse, complex, create, perform, set, friendly, experimental, user, type, collaborative, case, require, exploit, interact
SG14	16	resource, gateway, access, user, community, xsede, provide, science, job, create, environment, enable, set, large, cipre, campus, csg, easy, 'apache-airavata', phylogenetic, sequence, run, plan, project, highly, power, growth, issue, 'engineering discovery', usage

modelling and simulation; (*ii*) SG9, SG10 and SG14 are focusing on the user support part (with terms including gateway and portal); (*iii*) SG6, SG11 and SG13 seems focusing on the systemic part (with terms including platform, infrastructure, grid, framework). By analysing the terms characterising each topic we observe that: (*a*) the high frequency of words like 'datum', 'analysis', 'process', and 'storage' in two Science Gateway topics (SG1 and SG4) stress the data-oriented nature of some studies; (*b*) SG3 highlights the geospatial-oriented nature of some studies (e.g. [13]); (*c*) some terms referring to "systems" are emerging, e.g. HUBzero [14] in SG3, MoSGrid [15] and SEAGrid [16] in SG5, SciGaP [17] in SG7, Apache Airavata [18] in SG14; (*d*) in SG12 we recognised education-based terms that show that science gateways can be used as education platforms.

Regarding VRE topics (Tab. III), VRE2, VRE6, and VRE7 turn out having a high degree of relatedness. By analysing the terms characterising each topic we observe that: (*a*) collaboration-oriented terms are in VRE1 and VRE2; (*b*) some topics are characterised by domain specific terms, e.g. VRE2 has medical terms like 'cancer', 'biomedical', 'clinical', VRE7 has cultural heritage terms like 'humanity', 'art', 'text', 'historical', and VRE8 has earth related terms like 'climate', 'climate change', and 'climatic'; (*c*) VRE5 is data-oriented with terms like 'datum', 'data', and 'dataset'.

Regarding VLab topics (Tab. V), a certain degree of relatedness was emerging for: VLab1 and VLab2, highlighting simulation oriented aspects; VLab3 and VLab 6, stressing the remote laboratory aspects; VLab4 and VLab12, combining education-oriented aspects with an application domain; and, VLab10, VLab13, and VLab14 dealing with robotic and mechanical facets. By analysing the terms characterising each topic we observe that: (*a*) terms about education appear across many topics (VLab8, VLab9, VLab10, VLab12, VLab14) indicate that studies on VLabs are mostly used in educational domains; (*b*) terms related with real/reality appear in many topics, e.g. VLab1, VLab6, VLab9, VLab10, VLab11, and VLab 13; (*c*) in VLab9 topic, we can observe the terms 'virtual reality', 'VR' and 'education' suggesting that many studies are combining them, e.g. in [19] a virtual reality application for teaching physics to high school learners is described; (*d*) the most frequent terms related to the disciplines are 'physics', 'chemistry', 'electronics', 'electrical', 'mechanical', 'robotic' and 'chemical'.

TABLE IV
TOPICS CHARACTERISING VIRTUAL RESEARCH ENVIRONMENTS SELECTED STUDIES

Topic	# Studies	Top 30 representative terms
VRE1	14	virtual, information, environment, study, work, scientific, collaborative, develop, time, community, offer, library, make, tool, knowledge, share, evaluation, include, online, food, report, collection, feature, result, base, database, source, record, evaluate, communication
VRE2	17	research, virtual, environment, support, collaboration, paper, provenance, laboratory, development, integrate, international, set, challenge, type, describe, requirement, project, fund, cancer, range, biomedical, establish, clinical, collection, registry, target, system, major, capability, european
VRE3	24	infrastructure, service, resource, provide, tool, researcher, create, describe, image, access, domain, environmental, solution, include, experience, language, repository, biodiversity, metadata, ecosystem, conduct, grid, share, implement, aim, ieee, cloud, make, computing, issue
VRE4	26	datum, system, user, analysis, web, information, base, architecture, provide, support, distribute, approach, visualization, reference, big, paper, access, address, portal, analytic, storage, dynamic, interface, challenge, build, exist, propose, heterogeneous, task, springer international
VRE5	22	datum, workflow, science, computer, data, software, application, open, platform, large, enable, facilitate, process, processing, complex, publish, result, dataset, analysis, scale, author, method, compute, biology, set, increase, engineering, policy, computational, develop
VRE6	10	research, vre, environment, management, design, virtual, datum, researcher, process, case, practice, specific, article, component, group, model, part, focus, implementation, activity, identify, purpose, application, discipline, demonstrate, cycle, introduce, conceptual, ojax, test
VRE7	22	project, digital, research, environment, humanity, technology, virtual, base, semantic, approach, web, reconstruction, model, link, art, integration, structure, interactive, enhance, visualization, text, general, interaction, improve, light, aspect, standard, ontology, historical, object
VRE8	17	science, environment, present, model, gateway, climate, business, require, impact, social, development, change, scientist, discuss, problem, future, aim, simulation, context, apply, climate change, climatic, regional, current, global, spatial, result, methodology, relate, framework

TABLE V
TOPICS CHARACTERISING VIRTUAL LABORATORIES SELECTED STUDIES

Topic	# Studies	Top 30 representative terms
VLab1	39	model, simulation, design, structure, simulate, base, circuit, approach, tool, test, dynamic, require, order, change, modeling, include, theoretical, build, behavior, lead, condition, code, analyze, realistic, potential, combine, logic, state, agent, computational
VLab2	35	datum, research, analysis, information, support, set, management, time, database, multi, provide, environment, workflow, data, processing, make, include, scientific, term, impact, researcher, input-output, building, environmental, technical, enable, produce, specific, author, international
VLab3	83	laboratory, remote, web, system, base, internet, ieee, access, electronic, labview, technology, device, server, provide, interface, user, time, remotely, hardware, set, instrument, instrumentation, connection, make, give, client, transfer, browser, iot, engineer
VLab4	42	method, material, result, power, study, module, energy, obtain, analysis, measurement, increase, medium, industry, present, important, develop, cost, high, flow, good, complex, drive, motor, apply, product, ieee, expensive, department, generation, key
VLab5	58	resource, platform, service, cloud, project, science, infrastructure, base, framework, architecture, collaborative, paper, share, grid, provide, flexible, describe, integration, future, build, enable, community, compute, distribute, research, focus, access, solution, collaboration, high
VLab6	68	system, control, process, time, design, controller, virtual, real, industrial, simulation, paper, automatic, present, algorithm, plant, physical, automation, parameter, level, implement, performance, advanced, technique, temperature, nonlinear, motion, thermal, strategy, input, include
VLab7	49	student, laboratory, engineering, result, virtual, study, test, teach, undergraduate, experience, class, evaluation, performance, evaluate, program, group, assessment, academic, conduct, approach, carry, year, chemical, high, institution, face, challenge, feedback, aim, develop
VLab8	37	application, tool, software, develop, user, development, educational, digital, interface, visualization, present, field, feature, provide, create, language, level, communication, easy, give, main, case, method, programming, functionality, package, computer, advanced, domain, graphical
VLab9	55	virtual, teaching, reality, interactive, education, technology, training, physics, teach, physical, world, study, quality, create, field, vr, technique, enhance, high, immersive, order, great, visual, area, discuss, construction, interaction, construct, prototype, show
VLab10	50	lab, virtual, laboratory, education, environment, online, present, paper, component, robot, distance, problem, hand, practice, robotic, real, develop, mobile, solution, programming, open, perform, experimentation, implement, make, simulator, space, fundamental, alternative, offer
VLab11	61	experiment, network, computer, virtual, design, platform, technology, experimental, base, laboratory, paper, implement, implementation, security, provide, ieee, problem, hardware, real, traditional, scenario, virtualization, show, conduct, perform, create, result, complex, task, solve
VLab12	70	learn, student, learning, environment, concept, education, activity, base, skill, practical, knowledge, improve, teacher, chemistry, experience, support, understand, learner, provide, object, game, theory, content, school, approach, classroom, online, effective, communication, show
VLab13	32	virtual, laboratory, science, development, real, equipment, propose, user, operation, process, introduce, apply, function, human, research, instrument, improve, possibility, interaction, procedure, aim, practice, ability, efficiency, characteristic, environment, finally, achieve, basic, testing
VLab14	63	laboratory, virtual, engineering, work, machine, describe, implementation, electrical, educational, development, university, information, paper, part, process, practical, article, discipline, device, engine, due, demonstrate, current, creation, unity, measure, mechanical, create, modern, equipment

TABLE VI
TOPICS CHARACTERISING ALL SELECTED STUDIES IN A SINGLE CORPUS

Topic	# Studies			Top 30 representative terms
	SG	VLab	VRE	
A1	36	26	117	datum, research, environment, analysis, support, information, visualization, project, collaborative, digital, researcher, data, collaboration, provide, open, management, image, share, vre, approach, big, challenge, make, describe, source, focus, dataset, repository, include, social
A2	1	201	0	student, virtual, learn, laboratory, engineering, education, learning, teach, experience, environment, reality, practical, study, concept, university, educational, understand, online, physics, activity, teaching, distance, practice, improve, hand, high, skill, result, evaluation, world
A3	2	129	1	system, simulation, control, process, present, time, interactive, model, design, paper, develop, tool, software, environment, virtual, level, simulate, dynamic, controller, robot, order, base, create, industrial, propose, implement, real, give, form, parameter
A4	59	18	7	application, user, web, base, tool, workflow, scientific, interface, software, framework, architecture, paper, portal, integrate, integration, set, build, requirement, support, create, specific, component, complex, enable, case, code, domain, feature, execution, analysis
A5	1	173	0	virtual, laboratory, experiment, design, lab, remote, base, platform, network, technology, paper, provide, real, implementation, experimental, ieee, system, computer, online, operation, present, propose, education, implement, traditional, describe, achieve, result, teaching, show
A6	1	122	0	laboratory, virtual, development, computer, work, system, technology, develop, internet, machine, equipment, device, problem, make, communication, electrical, hardware, include, electronic, field, test, power, labview, high, ieee, software, process, instrument, information, cost
A7	14	54	8	model, study, result, method, structure, develop, tool, require, present, process, material, area, development, impact, energy, approach, analysis, potential, obtain, increase, case, scale, field, discuss, quality, detail, number, apply, order, demonstrate
A8	191	17	9	science, gateway, resource, service, infrastructure, cloud, access, community, provide, compute, computing, computational, grid, distribute, user, datum, management, platform, enable, capability, project, large, describe, storage, high performance, run, scientist, challenge, hpc, xsede

The similarities between SG and VRE topic distributions were expected to some extent. Like the VRE topics, that show two main VRE classes, namely data- and workflow-oriented, we could notice a progressive specialisation in the distribution of the SG topics, so that the overlaps are mainly limited to the different community-related practices and needs, with the exception of the first, fourth and the thirteenth topics, that identify two of the three specialisations of the SG classes, the first data-oriented, the latter workflow-oriented. While being a minor subclass, since it consists of 6% of the SG-related studies, the third SG subclass identifies an education-oriented typology, which shares characteristics with the education-oriented VLab one. The VLab topic distribution is of particular relevance for our study, since it helps identify two main "families" of VLabs: one education-oriented, with characteristics that are more suited to a training environment, the other research-oriented, which is in line with the previously observed SG and VRE subdivisions. The second one is further specialised into three classes, the first data-oriented, the second simulation-oriented, and the third focused on laboratories that are remotely accessible. The difference between the two VLab subclasses, the education- and the research-oriented, is attested by the delimiting characteristics that can be found in each one of the two topic groups. In the first group these characteristics are chiefly distributed in the two sets (*i*) 'student', 'education', 'teach', 'learn'; and (*ii*) ('virtual') 'reality', 'immersive', 'interactive', so that this group might be better defined by the absence of other properties (e.g. there is no reference to 'data'). In the second group, they encompass properties previously seen in the SG and VRE topics, for instance, 'data', 'research', and 'analysis'.

Topic modelling was also used to analyse the corpus as a whole. Table VI documents the 8 resulting topics. A certain degree of relatedness emerged regarding topics A1 and A8, and, A5 and A6. While A5 and A6 are dominated by VLab studies, A1 is dominated by VRE studies and A8 is dominated by SG studies thus suggesting an interconnection between the two families of studies. Since VLab studies represent the vast majority of our corpus, topics A2, A3, A5, A6, and A7 are dominated by Virtual Laboratory studies. Topic A2 contains terms like 'student', 'learn', 'teach', 'education' and 'study', indicating that this topic is education-oriented. Topic A3 terms distribution highlights the simulation-oriented nature of VLab studies in the field of engineering and design.

D. Threats to validity

Studies like this are vulnerable to threats of validity [20].

The most notable limitations of this study can be summarised in the following points: (*i*) the time range was restricted from 2010 onward; (*ii*) we considered only journal publications and conference proceedings; (*iii*) we did not perform any snowballing; (*iv*) we excluded the entries without a DOI; (*v*) we relied on titles and abstracts for the manual review; (*vi*) topics relatedness was based on a global view of each topic model obtained by multidimensional scaling. The main reason for the year limitation concerns the relevance of the dataset. While an extended time interval might seem more appealing, we observed that before 2010 the distribution of the retrieved publications per year did not justify its inclusion since its mean (μ=51.5) is significantly lower than the one of the distribution from 2010 on (μ=217.4) hence the 2010 year restriction represented a reasonable choice for a low interval limit. Because of the methodology adopted and the extent of the results, with an initial dataset consisting of 7775 entries (6037 VLab related, 1215 SG related, and 523 VRE related), we considered our corpus to be representative of the inquired topic and suitable for obtaining meaningful results. As a consequence we did not apply any snowballing technique to our search. Many entries (namely 1450) were removed because of the lack of a DOI. While this decision is certainly drastic, we believe it was a non prejudicial and a necessary one. It was non prejudicial because the resources we were interested in (conference papers and journal articles in the main literature) typically have DOIs assigned, especially if we take into consideration those published after 2010. It was necessary since the excessive number of identification issues would have undermined the validity of the results. Given the five different data sources, the variety of the retrieved metadata was a major factor in our decision. The manual process of revision is based on titles and abstracts rather than the full text of the papers. As a matter of fact, the extent of the corpus made it difficult to assess the relevance of each paper on a full-text basis. We decided to refer to the full text only for integrating what we considered missing, incomplete or ambiguous among the information provided by titles and abstracts. The last threat to the validity of this study is that we based our topic analysis mainly on two factors: the terms characterising each topic and the visualisation of the topics obtained by using LDAvis [21]. The two-dimensional visualisation provided by LDAvis is created through multidimensional scaling [22], which is admittedly one of the possible solutions, providing just one of the possible dimensionality reduction. Still by relying on both keywords and visualisation we could eventually mitigate any distortion.

V. Conclusion

This study aimed to systematically analyse the literature published in the last 12 years on Virtual Research Environments, Science Gateways and Virtual Laboratories. These three terms are used to describe a wide and varied class of systems and solutions proposed in diverse contexts to support the needs of specific communities.

The preliminary analysis was driven by three research questions: (*i*) what is the distribution over time of the literature on these systems and solutions? (*ii*) where the selected studies have been published? (*iii*) what are the major topics characterising these studies? Moreover, are the three terms actually referring to similar things or are they alluding to any peculiarity?

From this analysis it emerged that: (*a*) studies on VLab (62%) exceeded studies on SGs (25%) and VREs (10%) with the highest number of studies per year on VLabs published recently; (*b*) a very limited amount of studies (1%) were characterised by keywords recalling more than one term among SG, VRE, and VLab; (*c*) the publication venues for studies on SGs and VREs were in common while VLabs studies were oriented to education-related venues; (*d*) studies on VLabs have a stronger relationship with real laboratories than SGs and VREs studies; (*e*) studies on VLabs lead to particular topics (4 out of 8) that are blindly shared with SGs and VREs studies.

These preliminary findings actually suggest further research questions to be responded to develop a better understanding of the research domain, e.g. education-oriented solutions represent a significant part of the studies; what are the typologies and peculiarities of these solutions with respect to the rest of the studies? How many diverse typologies of systems exist, what are the peculiarities of each and the relative diffusion? Collaboration is a characteristic of the definitions of all the three terms, yet it emerged only in VREs, so what are the typologies of collaborations the diverse studies promote? What are the most frequent solutions and technologies exploited to develop these systems and solutions? How diverse is the lifetime of the various solutions and systems?

Data availability

The data that support the findings of this study are openly available on Zenodo at http://doi.org/10.5281/zenodo.6481183 [23].

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Blue Cloud project (grant agreement No. 862409), DESIRA project (grant agreement No. 818194), EOSC-Pillar project (grant agreement No. 857650), and SoBigData-PlusPlus (grant agreement No. 871042).

Author Contributions

According to CRediT taxonomy, authors contributed as follows: MA, AG, and DM performed Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization; LC performed Conceptualization, Methodology, Writing - Review & Editing, Supervision, and Funding acquisition; DC and PP performed Conceptualization, Supervision, and Funding acquisition.

References

[1] K. A. Lawrence, M. Zentner, N. Wilkins-Diehr, J. A. Wernert, M. Pierce, S. Marru, and S. Michael, "Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 16, pp. 4252–4268, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3526

[2] L. Candela, D. Castelli, and P. Pagano, "Virtual research environments: an overview and a research agenda," *Data Science Journal*, vol. 12, pp. GRDI75–GRDI81, 2013.

[3] M. Barker, S. D. Olabarriaga, N. Wilkins-Diehr, S. Gesing, D. S. Katz, S. Shahand, S. Henwood, T. Glatard, K. Jeffery, B. Corrie, A. Treloar, H. Glaves, L. Wyborn, N. P. C. Hong, and A. Costa, "The global impact of science gateways, virtual research environments and virtual laboratories," *Future Generation Computer Systems*, vol. 95, pp. 240 – 248, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X18314018

[4] P. Calyam, N. Wilkins-Diehr, M. Miller, E. H. Brookes, R. Arora, A. Chourasia, D. M. Jennewein, V. Nandigam, M. Drew LaMar, S. B. Cleveland, G. Newman, S. Wang, I. Zaslavsky, M. A. Cianfrocco, K. Ellett, D. Tarboton, K. G. Jeffery, Z. Zhao, J. González-Aranda, M. J. Perri, G. Tucker, L. Candela, T. Kiss, and S. Gesing, "Measuring success for a future vision: Defining impact in science gateways/virtual research environments," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 19, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6099

[5] L. E. Sepúlveda-Rodríguez, J. Garrido, J. C. Chavarro-Porras, J. A. Sanabria-Ordoñez, C. A. Candela-Uribe, C. Rodríguez-Domínguez, and G. Guerrero-Contreras, "Study-based systematic mapping analysis of cloud technologies for leveraging it resource and service management: The case study of the science gateway approach," *Journal of Grid Computing*, vol. 19, no. 4, p. 41, 2021. [Online]. Available: https://doi.org/10.1007/s10723-021-09587-7

[6] A. Diwakar, S. Poojary, and S. B. Noronha, "Virtual labs in engineering education: Implementation using free and open source resources," in *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, 2012, pp. 1–4.

[7] O. Panasiuk, L. Akimova, O. Kuznietsova, and I. Panasiuk, "Virtual laboratories for engineering education," in *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)*, 2021, pp. 637–641.

[8] B. Kitchenham, "Procedures for performing systematic reviews," Keele University, Technical report TR/SE-0401, 2004.

[9] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE'08. Swindon, GBR: BCS Learning & Development Ltd., 2008, pp. 68–77.

[10] S. Shahand, A. H. van Kampen, and S. D. Olabarriaga, "Science gateway canvas: A business reference model for science gateways," in *Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models*, ser. SCREAM '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 45–52. [Online]. Available: https://doi.org/10.1145/2753524.2753527

[11] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437920300703

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://www.jmlr.org/papers/v3/blei03a.html

[13] Y. Liu, A. Padmanabhan, and S. Wang, "Cybergis gateway for enabling data-rich geospatial research and education," in *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, 2013, pp. 1–3.

[14] M. McLennan and R. Kennell, "HUBzero: A platform for dissemination and collaboration in computational science and engineering," *Computing in Science & Engineering*, vol. 12, no. 2, pp. 48–53, 2010.

[15] J. Krüger, R. Grunzke, S. Gesing, S. Breuers, A. Brinkmann, L. de la Garza, O. Kohlbacher, M. Kruse, W. E. Nagel, L. Packschies, R. Müller-Pfefferkorn, P. Schäfer, C. Schärfe, T. Steinke, T. Schlemmer, K. D. Warzecha, A. Zink, and S. Herres-Pawlis, "The mosgrid science gateway – a complete solution for molecular simulations," *Journal of Chemical Theory and Computation*, vol. 10, no. 6, pp. 2232–2245, 2014, pMID: 26580747. [Online]. Available: https://doi.org/10.1021/ct500159h

[16] S. Nakandala, S. Pamidighantam, S. Yodage, N. Doshi, E. Abeysinghe, C. P. Kankanamalage, S. Marru, and M. Pierce, "Anatomy of the seagrid science gateway," in *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, ser. XSEDE16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2949550.2949591

[17] M. Pierce, S. Marru, E. Abeysinghe, S. Pamidighantam, M. Christie, and D. Wannipurage, "Supporting science gateways using apache airavata and scigap services," in *Proceedings of the Practice and Experience on Advanced Research Computing*, ser. PEARC '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3219104.3229240

[18] M. E. Pierce, S. Marru, L. Gunathilake, D. K. Wijeratne, R. Singh, C. Wimalasena, S. Ratnayaka, and S. Pamidighantam, "Apache airavata: design and directions of a science gateway framework," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 16, pp. 4282–4291, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3534

[19] N. P. Pandeka, P. A. Owolawi, T. Mapayi, V. Malele, G. Aiyetoro, and J. S. Ojo, "Mobile virtual reality (vr) for science projects: Ohm's law laboratory," in *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2021, pp. 01–06.

[20] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," *Information and Software Technology*, vol. 106, pp. 201–230, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584918302106

[21] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 63–70. [Online]. Available: https://aclanthology.org/W14-3110

[22] J. Chuang, D. Ramage, C. Manning, and J. Heer, "Interpretation and trust: Designing model-driven visualizations for text analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 443–452. [Online]. Available: https://doi.org/10.1145/2207676.2207738

[23] M. Arezoumandan, L. Candela, D. Castelli, A. Ghannadrad, D. Mangione, and P. Pagano, "Virtual research environments ethnography: a preliminary study," Zenodo, Dataset, 2022. [Online]. Available: http://doi.org/10.5281/zenodo.6481183