CYBERINFRASTRUCTURE REQUIREMENTS FOR RESEARCH COMMUNITIES

Tayane F. Peixoto[1], Luiz A. Martinazzo[2], Rosina O. Weber[3]

Abstract. *Cyberinfrastructure refers to the computational infrastructure that supports the productivity and impact of scientific fields. Many scientific fields of study rely on science gateways such as HUBzero to build cyberinfrastructure portals that offer standard features such as databases, simulation, visualization, and grid tools. This article aims to analyze requirements that align with the main goals of cyberinfrastructure. Some of the goals are to promote research, encourage interdisciplinarity, automate repetitive research steps, and to enable reproducibility and data sharing, among others, thus supporting the productivity of expenditures in science & technology. We analyze the problem cyberinfrastructure is intended to solve by focusing on entry barriers to research communities, illustrating it with the case-based reasoning community. Our study suggests additional features to existing cyberinfrastructure portals such as services to help researchers find funding and job opportunities, and where employers and funders can find researchers (e.g., expert locator engines), and the inclusion of a citation management system to facilitate creation and management of metadata for scientific workflows and other products.*

Keywords: *cyberinfrastructure, requirements engineering, root cause analysis, science gateways.*

[1] School of Engineering – Industrial Engineering & Management Systems Department – Fluminense Federal University (UFF) Niterói – RJ – Brazil. Email: tfogacia@id.uff.br
[2] Industrial Engineering & Management Systems – Federal University of Pelotas (UFPel) Pelotas – RS – Brazil. Email: lamartinazzo@gmail.com
[3] College of Computing & Informatics – Drexel University Philadelphia – PA – USA. Email: rosina@drexel.edu

1 INTRODUCTION

Cyberinfrastructure (CI) (Atkins 2003) is a vision of using computational resources in support of academic research, which may be interpreted as the infrastructure for the related concept of e-science (Hey 2001). The definition of e-science is broader and open-ended, aiming at effective global collaboration between groups of scientists. The overall goals of CI are to promote collaboration and interdisciplinarity (e.g., Gil *et al.* 2014). Despite the wide adoption of science gateways for CI such as HUBZero (McLeennan & Kennell 2010), we found no alignment with the commonly offered features with the main goals of CI. In this paper, we aim to bridge this gap by analyzing requirements for CI to identify features that align with CI. The methodology we adopt in this work is root cause analysis for requirements engineering (RE) (Leffingwell & Widrig 2006). The root cause analysis strategy (Andersen & Fagerhaug 2006) includes the steps of understanding the problem, exploring its possible causes via brainstorming, collecting and analyzing data, identifying the root causes, and suggesting ways to eliminate them. Our analysis identified the problem of entry barriers as the main problem that CI needs to solve to fulfill its goals. Entry barriers are obstacles that hinder collaboration and interdisciplinarity in research communities. The resulting recommended features consequently focus on eliminating those barriers.

Although the vision of CI and e-science is to serve an audience beyond researchers, including, for example, students, practitioners, citizen scientists, industry, and potentially the public in general, researchers from all fields (i.e., sciences, engineering, humanities) are the main users of CI tools and services. We hence focus on researchers as the users of CI. In this paper, thus, we refer to users as researchers. We illustrate our analysis with the case-based reasoning (CBR) research community. This way, users would be anyone interested in doing research in CBR. We contend that the analysis applied to other related research fields with similar characteristics should lead to very similar results.

Section 2 provides background for this work. Section 3 describes the implementation of the root cause analysis for RE. Section 4 presents the analysis and conclusions. The conclusions are organized in categories with recommended features for implementations that share CI goals such as science gateways. We conclude in Section 5.

2 BACKGROUND

2.1 CYBERINFRASTRUCTURE

CI can be described as an infrastructure based upon distributed computing, data, information, knowledge, and communication technology services (Atkins 2003). The myriad of technologies available today are drivers for CI systems in support of research. Some of these technologies are networked ontologies (Suárez-Figueroa *et al.* 2012), grid computing (Foster & Kesselman 2003), scientific workflows (Barker & Van Hemert 2007, Gil et al. 2007), linked data (Heath & Bizer 2011), service-oriented computing (Foster 2005), and cloud computing (Dillon *et al.* 2010). An excellent example of the goal of CI can be found at http://www.nsf.gov/news/special_reports/cyber/. In other words, CI is an infrastructure for knowledge, and, as outlined by Atkins (2003), just like physical infrastructure were required for the industrial economy, CI is required for the knowledge economy.

An important driver for CI is the need to promote interdisciplinary research given the need to address societal problems that requires expertise from multiple disciplines. CI needs to support and motivate research projects that promote novel contributions that solve complex social problems such as poverty, pollution, and sustainability.

Science gateways, defined as "*community-developed set of tools, applications, and data that is integrated via portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a target community*" (https://www.xsede.org/gateways-overview), are an important component of CI. It is common that communities use standardized science gateways such as HUBZero and each community uses a different set of requirements. As suggested by Gannon *et al.* (2007), five components are common to the science gateways: Data search and discovery; security; private data storage; tools for designing and conducting computational analysis, that is, "workflow" tools; and data provenance tracking.

2.2 REQUIREMENTS ENGINEERING

RE is the process of defining, organizing, and documenting requirements for system engineering and software engineering, RE's goals are to guarantee that the system's features and requirements fulfill all stakeholders' needs (Leffingwell & Widrig 2006), that the software addresses what it is intended to address, ad that the project succeeds (Nuseibeh & Easterbrook, 2000). It is therefore crucial that the RE team understands the problem or opportunity the target software is intended to address. The RE team should interact with all stakeholders to assess whether the comprehension of the problem or opportunity is consensual.

One crucial step in problem analysis is root-cause analysis. Root-cause analysis helps the team comprehend the problem and its contexts within the universe of its users and other stakeholders so that they can propose useful features. The RE method we rely on proposes a table format to state the problem (Leffingwell & Widrig 2006). The table has four fields that make explicit the main aspects of the problem or problems. The fields are the problem statement, stakeholders that are affected by the problem, the consequence that negatively impacts each stakeholder group, and the benefits of a solution. Our experience applying this method suggests these benefits should be completely divorced from one or another specific solution. This representation is illustrated in Table 2, in Section 3.

2.3 ROOT-CAUSE ANALYSIS

A problem represents a challenge that inspires a solution that provides circumstances that are more desirable (Andersen & Fagerhaug 2006). Problems are events that may lead to undesirable consequences or prevent an expected positive result. The main goal of the root cause analysis method is to identify the bottom cause or causes for a problem and to define a solution that will eliminate or prevent it. Root-cause analysis can be described as an investigation organized in defined steps that aim to find the main cause of a specific problem and to delineate actions that can eliminate it (Andersen & Fagerhaug 2006). According to Gano (2007), to ensure its effectiveness, root-cause analysis should have a clear description of how the proposed solutions meet the goal to eliminate the source of the problem so it does not recur. The analysis should clearly define the problem and its significance to the problem owners (i.e., the ones affected by it). The analysis should delineate the known causal relationships that combined to cause the problem. The analysis should also establish causal relationships between the root cause(s) and the defined problem. It should present the evidence to support the existence of identified causes of the problem. It should also explain how the solutions would prevent the recurrence of the defined problem. Finally, it should document all these criteria in a final root-cause analysis report so others can easily follow the analysis.

The cause of a problem can be classified in three categories. First-level causes, that are the causes that directly lead to the problem; higher-level causes, the ones that lead to the first-level causes; and compound causes, which are multiple factors that combine to cause the problem (Andersen & Fagerhaug 2006).

The analysis approach proposed by Andersen and Fagerhaug (2006) consists of seven steps: Problem Understanding, the first step, aims at understanding the nature of the problem. Problem Cause Brainstorming, the second, seeks to gather ideas of root causes. Problem Cause Data Collection, the third, relies on generic tools and techniques (e.g., interviewing) to systematically and efficiently collect data related to a problem and its probable cause. Problem Cause Data Analysis, the fourth, analyzes the data from different angles to assess whether there is evidence pointing to consistent or contradicting conclusions. Problem Cause Identification, the fifth, depends on the analysis of the causes to establish primitive problems. Root Cause Elimination, the sixth, devises solutions that remove the root cause and thus eliminate the problem or prevent its recurrence. The last step is Solution Implementation that varies depending on the nature of the problem under analysis. The tools and methods respective to each step are laid out in Table 1.

Table 1 – Problem Solving Process Tools and Methods

Step	**Tools**
Problem Understanding	Flowcharts, Critical Incident, Spider Chart, Performance Matrix
Problem Cause Brainstorming	Brainstorming, Brainwriting, Is-is not Matrix, Nominal Group Technique, Paired Comparisons
Problem Cause Data Collection	Sampling, Surveys, Check Sheets
Problem Cause Data Analysis	Histogram, Pareto Chart, Scatter Chart, Problem Concentration Diagram, Relations Diagram, Affinity Diagram
Problem Cause Identification	Cause-and-Effect Charts (i.e., Fish-bone and Process Chart), Matrix Diagram, Five Why´s, Fault Tree Analysis
Root Cause Elimination	Six Thing Hats, Theory of Inventive Problem Solving (TRIZ), Systematic Inventive Thinking (SIT)
Solution Implementation	Tree Diagram, Force-Field Analysis

Source: Andersen and Fagerhaug (2006)

2.4 CASE-BASED REASONING

Case-Based Reasoning (CBR) (Richter & Weber 2013) is a reasoning methodology that relies on reusing previous solutions to solve new problems. CBR is usually studied in computer science and related disciplines, but it has vast applicability in many other engineering, medical, and social science applications. This community has annual events since late eighties. As a field of study, it is closely related to cognitive science, machine learning, knowledge-based systems, and recommender systems.

3 ROOT-CAUSE ANALYSIS IMPLEMENTATION

This section covers the three first steps of the method presented in Table 1, namely, Problem Understanding, Problem Cause Brainstorming, Problem Cause Data Collection. The problem we address in this implementation is the entry barriers imposed on researchers who want to enter or investigate a field of study that we here illustrate with CBR. Next, we describe steps we undertook (Andersen & Fagerhaug 2006).

3.1 PROBLEM UNDERSTANDING

The magnitude of a problem like entry barriers into a field of study requires study of the activities researchers undertake, and an introduction to the target field. This expertise was available within these authors and therefore this step was of internal knowledge transfer.

3.2 PROBLEM CAUSE: BRAINSTORMING

The problem understanding step leads to the problem statement (Leffingwell & Widrig 2006). The problem statement format ensures the problem is well-defined and the solution benefits are clear. The problem statement for our target problem is given in Table 2, where the problem identified refers to the barriers faced by researchers to study the target field. The stakeholders affected by it are newcomers and researchers already in the field. This is because existing barriers could potentially cause established researchers to abandon the field.

The negative results are caused by the barriers, which prevent stakeholders from engaging in the field. The negative results prevent researchers to conduct research in the field, obstructing achievement of CI goals. This is how we identify results such as lack of data, lack of opportunities to collaborate, lack of funding, lack of jobs, etc. The benefits of a solution should neutralize negative results. They are consistent with a vision of reality of a field without barriers, or where there are incentives.

Table 2 – Problem Statement

Element	Tools
The problem the organization faces is that	… there are entry barriers for researchers to study a subject or field of study (directly indicated which community)
Which stakeholder groups are affected by this problem?	Researchers (established and newcomers)
What is (are) the negative result(s) affecting each of the stakeholder groups mentioned above?	Insufficient interest in community, lack of interdisciplinary projects, lack of collaboration, lack of significant impact, limited knowledge of the field, lack of jobs in the field, lack of classes being taught, lack of funding opportunities, lack of data, lack of tools, difficulty to write literature reviews, etc.; it is hard to learn about the field.
What characterizes a solution that could solve this problem?	Researchers easily determine whether they want to study the subject; there is a focused source of knowledge about the field (e.g., for writing literature reviews); researchers can learn about the field using existing tools and data.

Source: Leffingwell and Widrig (2006).

3.3 PROBLEM CAUSE: DATA COLLECTION

We started this step by implementing it without external input. The results of this is a list of potential root causes for the problem. We refer to this as our initial list. The main approach we used for data collection was interviews. In this subsection, we describe the protocol and implementation of the interviews. Interviews were complemented by web search, which we describe in the next section together with analysis and identification.

We conducted interviews with four CBR researchers using the following protocol. First, we asked what barriers they could identify to entry the CBR field. We took note of all the barriers the interviewees suggested. For each barrier, we examined whether there was a fit to one of the existing categories in our initial list or if a new category was needed. We clarified any potential misunderstandings at that point. Once interviewees exhausted root causes they could name and we clarified any further aspects, we then presented our initial list asking for their assessment. This step prompted further discussions and clarifications of the categories and the actual root causes. For example, under the category datasets, one interviewee mentioned that other fields have datasets associated to standard algorithms that represent a benchmark against which other researchers can test different algorithms. The final step was to ask the interviewees for last comments about barriers in the field or any problems they could think about the field even if not directly associated with a barrier. This was an attempt to reveal undiscovered ruins, as sometimes a general problem may not be obviously perceived as a barrier, but a relation is possible. Our goal was to explore all potential ideas. For example, one of the interviewees pointed out the number of program committee members and the lack of a steering committee in the conference organization as a possible problem. Between these interviews and the time we prepared the final version of this work, the community had added a steering committee to their conference organization.

4 ANALYSIS AND REQUIREMENTS

This section includes the three of the seven steps of the analysis approach previously described, which are Problem Cause Data Analysis, Problem Cause Identification, and Root Cause Elimination. As described in the previous section, we identified categories of root causes during the first stages. These categories suggest that the causes of the target problem can be classified as compound causes (Andersen & Fagerhaug 2006) (Section 2.3). The root causes preventing researchers to engage in the CBR field are organized in the categories: general resources, tools, datasets, journals, conferences, applications, jobs, teaching, funding, impact, and reproducibility.

In this section, we analyze each category of root causes, and propose a list of features that could eliminate the problem addressed. These features are recommended to comprise any CI implementation for the CBR community that aims to reduce entry barriers.

The organization and choice of labels for these categories is not rigorous. The first category is called general resources to distinguish from other categories that, despite possible also representing a type of resource, deserve a category of its own due to their importance, which is the case with the category Data.

Each category is associated with root causes of barriers for research in the CBR field. The general idea is that elements related and thus classified under each category can influence the entry of a newcomer to the field or limit in some way the work of an established researcher. The elements under each category should exist, be available, and be of good quality. This means that, in general, the lack or low quality of these elements may impose a barrier to researchers. Next, we describe our analysis of these elements that led to the identification of the root causes.

4.1 GENERAL RESOURCES

The general resources that we discuss in this paper are commonly found in general portals or wikis that scientific communities maintain (e.g., the CBRWiki). These are general resources such as list of research groups, authors, books, dissertations, and presentations. We grouped some of these under this label because these are trivial and do not require extensive analysis.

The resources we found available for the CBR community are: Applications, Books, Case Bases, CBR Readings, CBR Topics, Course Material, Industry, Information Sources, Jobs, Meetings, People, Presentations, Reproducibility, Research Groups and Tools. The analysis of these resources is very straightforward; they should be made available for ease access as they can provide valuable information for a background and introduction to the field.

4.2 TOOLS

We analyzed seven CBR tools. This step was carried out by the authors that had no knowledge of the CBR field prior to engaging in this work. These authors searched, downloaded, and attempted to use the tools; they reported on their steps, perceptions, actions, and reactions. This retrospective reporting helped us identify key features that may represent a significant barrier for newcomers. In Table 3, we show the analysis of the CBR tools with some selected parameters.

Table 3 – CBR Tools

	CAKE	**CBR Shell**	**CBR Works**	**COLIBRI Studio**	**eXiT CBR**	**Free CBR**	**my CBR**
OS	Windows	Windows	Windows	OS X, Windows	Windows	Windows	OS X, Windows
Programing Skills	Required	NA	Not required	Required	Not required	Not required	Not required
Easy to Use	No	NA	Yes	Reasonable	NA	Yes	Reasonable
Tutorial Available	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Free	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NA = No analysis was made as the tool was not used. Results of analysis conducted from May to August 2015

Our analysis concluded that some features are decisive for a newcomer to enter a new field. The features are that tools should: 1) Be accessible in CI systems; 2) Do not require that users know how to write code in any particular programming language; 3) Be easy to use, and also include tutorials so someone without previous knowledge can implement an example of the method; 4) Be compatible with most widely used and recent operating systems; 5) Be free and offer full-capability versions for research purposes; 6) Be publicized in central community repositories (e.g., wikis); 6) Incorporate features that exemplify key characteristics of the method under study, such as different similarity metrics for case retrieval (i.e., in case of CBR), and evaluation functions such as LOOCV.

4.3 DATASETS

Datasets are extremely important for researchers as they enable them to experience the methodology. They describe scenarios, exemplify uses of methods, and illustrate applicability to multiply domains. Datasets are fundamental to demonstrate validation of any contributions, making them indispensable when newcomers select dissertation topics, as we learned anecdotally from doctoral students.

The suggested features with respect to datasets comprise the existence, reliability, and usability of datasets in a research CI platform. These features are: 1) Quality search capability that searches datasets in all contents and metadata (e.g., textual dataset); 2) Capability to export (e.g., to workflows) and import data; 3) Capability to edit metadata; 4) Capabilities for merging and fusion with other datasets; 5) Enable association with publications; 6) Include benchmark datasets; 7) Metadata should include information about provenance, and accuracy.

4.4 JOURNALS

The digital search revealed articles related to CBR in 20 different journals. There is no main journal or group of journals addressed to the topic. This reveals the lack of a unified forum where newcomers can search for the latest developments in the field.

Researchers rely on scientific journals as a source of literature review, recent contributions, and open research questions. Having its body of knowledge spread in too many journals promotes an entry barrier because it makes it hard to conduct a reliable exhaustive literature review. As a source of reference, we compare the CBR with the field of economics, where the main research appears in a handful of first-tier journals (e.g., AER, JPE, QJE).

Although it is not addressed directly by a feature, this entry barrier may be attenuated by the construction of an adequate CI system. For the purposes of literature review, the features here proposed are also valid for conference papers, being them: 1) Capability to identify when a new article or paper is published; 2) Capability to gather metadata about articles or papers in the field; 3) Capability to search full text and metadata of articles and papers in the field; 4)

Capability to gather and make searchable research questions in the field; 5) Capability to collect and make searchable literature reviews in the field; and 6) Capability to associate research questions, literature reviews, datasets, and workflows.

Ultimately, the focus of these resources may help publishers assess the need for a dedicated journal. The features in this subsection are mainly about publications, so we also referred to conference papers. We address other aspects of conferences in the next sub-section.

4.5 CONFERENCES

The digital search revealed 24 conferences with published papers that mention CBR, though only one is specifically about CBR, the International Conference on Case-Based Reasoning (ICCBR). Our analysis of conferences did not reveal any direct features that could eliminate barriers. Analyzing conference management is outside the scope of this work.

4.6 APPLICATIONS

Commercial, industrial, or governmental applications provide an exogenous source of motivation for research questions in a field. Having information about those is considered important to researchers, in general, as they illustrate the relevance of the methodology. The entry barrier would be not having access to information about applications. Ultimately, following the very nature of CBR, access to existing applications may be used as examples for further application development.

The recommendation to eliminate this barrier is to include features to collect and manage a repository of applications. The suggested features for applications are analogous to the features for publications: 1) Capability to identify when a new application is developed or deployed; 2) Capability to gather metadata about applications; 3) Capability to search full text and metadata of applications; 4) Capability to collect and make searchable literature reviews related to applications; and 5) Capability to associate literature reviews, datasets, and workflows with applications.

4.7 JOBS

The availability of job opportunities is crucial for researchers, it can directly influence the motivation to pursue that field. Analogous to journals and applications, we recommend that a reliable repository with jobs in the field be maintained as a resource in the CI system. The most important benefit of a CI system is how it can help improve transparency of the field and consequently influence the offer of jobs. Consider how just the features listing applications will make the methodology better known and understood. Once organizations adopt it, the jobs will follow.

Specific to CBR, in our search, we found both academic and non-academic jobs. In this field, jobs that mention CBR may not require knowledge of the field because they may simply require people to use a CBR system. Recommended features are: 1) Identify when a new job is offered; 2) Job search; 3) Expert locator system; and 4) Capability that associates jobs with experts and that communicates experts about those associations.

4.8 TEACHING

Lack of courses teaching a methodology is an obvious entry barrier. Further analysis of interfering with this root cause falls outside the scope of RE. As in other categories, keeping up-to-date resources has the potential to increase transparency and improve the quality of existing courses with the dissemination of course materials.

The suggested features are: 1) Capability to identify when a new course if offered or substantially revised; 2) Capability to gather metadata about courses; 3) Course search; and 4) Course entry.

4.9 FUNDING

Funding is essential for research, so that the lesser the funding, the greater the barrier. As with jobs and teaching, the CI system can increase transparency of the field and ultimately influence the offer of funding opportunities. We do not again propose features that would have a direct influence on the offer of funding for a research field because this would require another type of analysis. We do recognize the importance of keeping researchers very well informed with an up-to-date resource about funding opportunities.

Besides this, the CI system can also provide features for program directors from funding organizations (i.e., private and governmental) in charge of composing calls for proposals. Some of these features are listed under the category Journal such as literature reviews and open research questions. The other suggested features are: 1) Capability to identify when a new funding opportunity is offered; 2) Capability to associate funding opportunities with experts and to communicate experts about those associations; 3) Capability to gather metadata about funding opportunities; 4) Funding search; and 5) Funding entry. Funders can benefit from such information to determine the target field in a call for proposals. Further gathering data from funders may be needed.

4.10 IMPACT AND REPRODUCIBILITY

The main feature proposed in the literature in support of a field's impact and its demonstration through reproducibility is scientific workflow management. Example features are 1) Workflow editor, 2) Search, 3) Reuse, 4) Adaptation, 5) Similarity, 6) Document drafting, and 6) Links to citation management. This is a CI module where CBR has been demonstrated as useful (e.g., Bergmann and Gil, 2014; Gil *et al.*, 2011). Additional features for workflow management are described in several publications such as those describing Taverna (Oinn *et. al.*, 2004), Pegasus (Deelman *et. al.*, 2005), Triana (Taylor *et al.*, 2015), etc.

In this aspect, the CBR community is ahead because of CAKE. CAKE is a tool that combines CBR and existing workflows by integrating knowledge and process management in a common platform (Bermann *et al.*, 2014).

4.11 OTHER

The categories dataset, journals, and applications include features to eliminate entry barriers that mention literature or publications. This imposes the need of tools to manage citations given the need to manipulate publication metadata. Given the vast offer of tools to manage citations, an existing tool or service could be incorporated to meet this requirement.

5 CONCLUDING REMARKS

This paper describes an analysis of entry barriers for researchers aiming to enter the CBR field. We used the CBR field to illustrate the analysis of root causes in order to propose features for CI implementations that eliminate entry barriers. The features listed in each category in Section 4 comprise the recommendations resulting from this work. Although they were conceived based on the analysis of one community, they are potentially useful to other research communities, mainly those communities that share similarities with the CBR field, such as the focus on a methodology, the use of data, and applicability to a variety of domains.

The features we recommend in Section 4 for CI systems are the main conclusions of this work. Some of these features are information requirements, such as applications, journals, conferences, and jobs; others are technical features, such as dataset and workflow support. Some are recommendations that may not even be implemented via software. The list of features is not exhaustive, but provides the main capabilities to address and neutralize the root causes for barriers to entry a field like CBR.

Existing CI systems lack some of these features, and therefore do not include the features that are aligned with promoting collaboration, interdisciplinarity, and research productivity.

which are the main goals of CI. Ideally, science gateways – shells for research CI systems, should therefore include features consistent with CI goals.

6 ACKNOWLEDGMENTS

The authors Tayane F. Peixoto and Luiz A. Martinazzo wish to thank the CAPES Foundation, Ministry of Education of Brazil, for its financial support during the program Science without Borders, which was crucial for this work.

REFERENCES

Aha, D. W. Naval Research Lab, Washington, DC. July 2015: private communication.

AIAI CBR Shell Aiai.ed.ac.uk, 2015. [Online]. [Accessed: 03- Jun- 2015] Available: http://www.aiai.ed.ac.uk/project/cbr/CBRDistrib.

Andersen, B., & Fagerhaug, T. (2006). *Root cause analysis: simplified tools and techniques*. ASQ Quality Press.

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*.

Barker, A., & Van Hemert, J. (2007, September). Scientific workflow: a survey and research directions. In *International Conference on Parallel Processing and Applied Mathematics* (pp. 746-753). Springer Berlin Heidelberg.

Bergmann, R., & Gil, Y. (2014). Similarity assessment and efficient retrieval of semantic workflows. *Information Systems*, *40*, 115-127.

Bergmann, R., Gessinger, S., Görg, S., & Müller, G. (2014, November). The collaborative agile knowledge engine CAKE. In *Proceedings of the 18th International Conference on Supporting Group Work* (pp. 281-284). ACM.

Cakeflow.wi2.uni-trier.de, 'CAKE | Home of the Collaborative Agile Knowledge Engine', 2015. [Online]. Available: http://cakeflow.wi2.uni-trier.de/welcome-want-a-cake/. [Accessed: 28- May- 2015].

Cbrwiki.fdi.ucm.es, "Cbrwiki", 2015. [Online]. Available: http://cbrwiki.fdi.ucm.es/mediawiki/index.php/Main_Page. [Accessed: 09- Jun- 2015].

Davis, D. (unpublished). *Where do I start.* Retrieved July, 2015 from http://www.columbia.edu/~drd28/Teaching.htm

Deelman, E., Singh, G., Su, M. H., Blythe, J., Gil, Y., Kesselman, C., ... & Laity, A. (2005). Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming*, *13*(3), 219-237.

Dillon, T., Wu, C., & Chang, E. (2010, April). Cloud computing: issues and challenges. In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on* (pp. 27-33). IEEE.

Exitcbr.udg.edu, 'CBReXiT- UdG', 2015. [Online]. Available: http://exitcbr.udg.edu. [Accessed: 03- Jun- 2015].

Foster, I. (2005). Service-oriented science. *Science*, *308*(5723), 814-817.

Foster, I., & Kesselman, C. (Eds.). (2003). *The Grid 2: Blueprint for a new computing infrastructure*. Elsevier.

Freecbr.sourceforge.net, 'FreeCBR', 2015. [Online]. Available: http://freecbr.sourceforge.net. [Accessed: 03- Jun- 2015].

Freecbr.sourceforge.net, 'FreeCBR', 2015. [Online]. Available: http://freecbr.sourceforge.net. [Accessed: 03- Jun- 2015].

Gaia.fdi.ucm.es, 'COLIBRI Studio | GAIA – Group of Artificial Intelligence Applications', 2015. [Online]. Available: http://gaia.fdi.ucm.es/research/colibri/colibristudio. [Accessed: 03- Jun- 2015].

Gannon, D., Plale, B., Christie, M., Huang, Y., Jensen, S., Liu, N., ... & Simmhan, Y. (2007). Building grid portals for e-science: A service oriented architecture. *High Performance Computing and Grids in Action*.

Gano, D. L. (2007). Comparison of common root cause analysis tools and methods. *Apollo Root Cause Analysis-A new way of thinking*.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., ... & Myers, J. (2007). Examining the challenges of scientific workflows. *IEEE Computer*, *40*(12), 24-32.

Gil, Y., Gonzalez-Calero, P. A., Kim, J., Moody, J., & Ratnakar, V. (2011). A semantic framework for automatic generation of computational workflows using distributed data and component catalogues. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*(4), 389-467.

Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science*, *346*(6206), 171-172.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, *1*(1), 2.

Hey, T. (2001, July). E-Science and the Grid. In *British National Conference on Databases* (pp. 23-23). Springer Berlin Heidelberg.

Leffingwell, D., & Widrig, D. (2006). *Managing software requirements: a unified approach*. Addison-Wesley Professional.

McLennan, M., & Kennell, R. (2010). HUBzero: a platform for dissemination and collaboration in computational science and engineering. *Computing in Science & Engineering*, *12*(2) (pp. 48-53).

Mycbr-project.net, 'myCBR', 2015. [Online]. Available: http://mycbr-project.net. [Accessed: 03- Jun- 2015].

Nuseibeh, B., & Easterbrook, S. (2000). Requirements Engineering: A Roadmap. In *International Conference on Software Engineering* (pp. 25-46), New York, NY, USA.

Ontañón, S. Drexel University, Philadelphia, PA. July 2015: private communication.

Richter, M. M., & Weber, R. (2013). *Case-based reasoning: a textbook*. Springer Science & Business Media.

Richter, M. Universitat Kaiserslautern. July 2015: private communication.

Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., & Gangemi, A. (Eds.). (2012). *Ontology engineering in a networked world*. Springer Science & Business Media.

Taylor, I., Shields, M., & Wang, I. (2015). Distributed p2p computing within triana: A galaxy visualization test case. In *IPDPS '03 Proceedings of the 17th International Symposium on Parallel and Distributed Processing* (pp. 16.1). Washington, DC, USA.

Xsede.org, 'XSEDE | Overview', 2015. [Online]. Available: https://www.xsede.org/gateways-overview. [Accessed: 28- Oct- 2015].