Rule-Based Querying of Distributed, Heterogeneous Data

Tom Lansdale, Peter Bloodsworth, Ashiq Anjum, Irfan Habib, Yasir Mehmood and Richard McClatchey

Centre for Complex Co-operative Systems, BIT, UWE Bristol, UK

Abstract

When searching for data, users tend to think in terms of the information they need to retrieve and not where and how it is stored. This is especially true in the highly complex domain of neurological research. The generic medical querying service described herein, aims to close the gap between users and data resources that are intricate, distributed and heterogeneous in nature. A theme that is common to both this domain and the emerging Internet of things is that users often need to query more than a single resource. This has come about with the large-scale fragmentation and distribution of data. Our work thus far has produced a prototype architecture for the querying of heterogeneous distributed data in the medical domain. A flexible querying mechanism that has the potential to support agent-based personalization has been developed to provide users with accurate and relevant results and some initial results are presented.

Keywords

Database, Data integration, Heterogeneous databases, Multi-agents systems, Querying, Rule-based.

1. Introduction

The Future Internet is likely to be much larger and indeed, more complex than the Internet as we currently know it [1]. With ever increasing data and more devices to interact with, information overload [2] could become an even greater problem than it is today. Data is also distributed in nature and growing in mobility. In addition to effective storage and database technologies, this will likely require efficient querying mechanisms to be put in place. These will need to assist users in accessing and making sense of all the information that will become available to them. Furthermore, it seems reasonable to assume that users will wish to query this "Internet of Things" using simple, perhaps even "Google-like" interfaces together with information, which is of value to them personally. The combination of simplified interfaces and more complex underlying data resources may bring the conclusion that a layer of intelligence is necessary to be translated between the user and the data environment.

Ontology has been widely used to contain domain knowledge and meta-data, which can be tailored to user's individual needs. They also play a key role in the storage of domain-specific information and often drive intelligent behavior [3]. In fact, some have called them 'germane to the idea of machine-processable data' [4]. They may therefore be employed to structure information regarding access methods to resources on the emerging Internet and allow this to be personalized, since not every user will have the same information retrieval preferences.

In this paper, we consider the similarities between the challenges that are posed by querying in the medical domain and in the emerging Internet. We explore how an approach that is being taken in the domain of neurological research may be more widely applied, and discuss the ways in which semantic meta-data can play a greater role in the structuring of the knowledge contained on the Internet. This work considers the adoption of semantic meta-data to represent and describe how information can be accessed and linked together. The focus being on how it can be harnessed to facilitate efficient querying, increased personalization and data integration. An architecture is presented, which facilitates the querying of heterogeneous, distributed data resources. We go on to describe the motivation for decisions made at the architectural design stage and outline the test-bed that has been implemented. Initial results are demonstrated thereafter and we describe how we envisage the work progressing in the future.

2. Similarities Between Neugrid and the Emerging Internet

The aim of the FP7-funded neuGRID project is to provide a user friendly grid-enabled e-infrastructure, which will enable the European neuroscience community to carry out research that is necessary for the study of degenerative brain diseases. This process involves parameters such as brain volume being extracted by applying image processing techniques on patient scans. These parameters are extracted from images by executing workflows or pipelines. Some pipelines generate approximately one thousand percent more data than they consume. It is clear therefore that querying will play a central role in the system by providing mechanisms for accessing the distributed data resources, which are inside the grid environment. The data is heterogeneous in nature and will

comprise provenance information made up of complex detail regarding executed pipelines as well as MRI and patient data. Local data can be specific to individual institutions and is therefore structured in different ways. Such data often comes in a variety of formats and it is challenging to query and integrate it.

The neuGRID architecture consists of an intelligent querying service, which has access to ontological domain data to process, enrich and personalize queries. It features a layer of abstraction allowing it to be connected to the kinds of distributed, heterogeneous data resources one can expect to find in such an object-rich environment as the Future Internet. A rule-engine has been used to harness this semantic data during the processing of queries. This could potentially drive personalized agents that act on behalf of users to interact with the next generation Internet. In doing so, we may also enable intelligent software to better access and exploit the data that is now becoming available on the Internet [5]. It is known that users are reluctant to changes in the way they interact with services. Simple interfaces have been proven successful by Google even as the data becomes increasingly complex and multi-modal. We propose that this may be achieved by building some intelligence into the underlying querying mechanism and keeping interfaces as static as possible.

3. Previous Work

The querying of distributed heterogeneous resources is becoming increasingly necessary as information continues to grow and become ever more complex. In response to such challenges, researchers have proposed a mixture of meta-data and abstraction layers to address the heterogeneities that are found within databases and other information sources. Such heterogeneities exist for many reasons, different vendors, underlying data structures and querying languages often being contributory factors. Whilst much progress has been made in this area, it is becoming clear that users often require more than basic access. They may need to work with heterogeneous information sources that are linked in some way. Therefore the integration and personalization of such data resources is of growing importance. One approach that has been widely employed is the use of a layer of abstraction, which acts as an interface to the underlying data resources and handles the differences between them. The concept of a FDBS (Federated Database System) was initially put forward by Hammer and McLeod (1979) [6]. Federated database systems (FDS) were described in greater detail by Sheth and Larson in1990 [7]. Here an attempt was made at merging physical databases into components of a larger seemingly virtual federated database.

However, a federational one is probably not enough to deliver both integrated and personalized querying due to a lack of semantic awareness. There is no additional knowledge about the user or about concepts within the domain a given resource describes. It is for this reason that additional semantic representation might be required to aid in the querying of such relational resources.

Today, several systems exist, which aim to allow the transparent querying of multiple, heterogeneous data resources. Google adopt advanced mechanisms to query the many types of documents which exist on the Internet, they also store semantic information to try to 'understand what users mean' when they type a query. It is arguable how well one can query data which is not well-represented semantically as is the case with most online content today. Users have to be increasingly more and more explicit when searching the Internet. They often feel that they are looking for an ever smaller needle in an ever growing haystack. It is for this reason that users may struggle in the future as not only the volume of information grows but also the types of information diversify.

A 'computational knowledge engine' named Wolfram Alpha [8] was recently released, which attempts to address the problem of statistical querying on the Internet. It is believed that without semantic meta-data, answers can only be found to questions, which have been literally asked before. Wolfram Alpha and Google can return relevant results based on a very simple input by applying complex natural language processing techniques. Wolfram have produced a means of semantically integrated querying but there is no element of personalization as yet, thus results are not individually tailored. Both Wolfram Alpha and Google provide a degree of intelligence to interpret the users' intentions, for example synonym awareness or spelling correction. However, personalization is not something that features explicitly in either of these services. The combination of personalization semantics and domain semantics offers the possibility of providing more relevant results and assisting users in dealing with complicated resources.

Personalization in querying has been investigated before by Bayardo *et al*. (1997) [9] who speaks of a user agent to act as a users' intelligent interface to a federated database. Indeed, Huhns recently spoke of agents for personalization at AAMAS 2009 [10]. It is believed that agents could be applied in this way on behalf of the user to query data in projects such as neuGRID, where individuals' specific data needs differ. Bayardo *et al*. present Infosleuth, a system to integrate heterogeneous information bases. In this paper, the authors accept that there is no control over the registration of new information sources and suggest the use of a user agent to access information based on concepts rather than keyword. The lack of control over

growing resources is the same in neuGRID and indeed the Internet. Any querying services built on top of these existing systems must be abstracted from the dynamic nature of the underlying resources. Using ontology to aid querying is not a new idea, indeed, Shah *et al*. (2008) use this approach in order to integrate data resources by concepts contained within ontology [11].

In summary, federation can bring resources together but may suffer from limited integration because semantic relationships between data are not captured. Interfaces for tools such as Google and Wolfram Alpha appear popular and usable but they only seem to work on a fairly limited set of data types, furthermore, these tools lack personalization. Agents can offer the promise of handling distributed, heterogeneous data [9] and acting intelligently to personalize information but often lack access to abstraction layers. Our aim is to create a generic and flexible architecture to allow the personalized querying of distributed, heterogeneous resources in neuGRID aided by semantic meta-data.

4. Querying Architecture

A querying service was sought to provide a means of accessing the data held in neuGRID. Data in this context is naturally distributed and heterogeneous, stored at institutions throughout Europe, therefore this data must be integrated. A generic architecture has been developed for the querying service, which focuses on linking semantic meta-data with a rule-engine to provide a highly customizable and flexible architecture. Being service oriented means that a standard interface is created, this is useful because it makes the development of client applications much easier. In an increasingly mobile world, this is an important consideration since it might not always be that users wishing to interact with the system are in an office.

A motivation for making semantics an integral part of this querying mechanism is that clinicians will use the service from different cultures and backgrounds. In the medical domain and neuroscience, in particular, there are many synonyms and preferences in the way aspects of research are described. It is the aim of the querying service to improve the users' experience when interacting with neuGRID platform through personalization that focuses on the needs of an individual clinical researcher. As well as written similarities between words, it might also be useful to store, for instance, the special relationships between structural features of the human brain. This may allow queries to be enhanced by possibly using such information to expand a basic query to include synonymous terms and related regions.

The querying service architecture is represented in Figure 1, it is made up of the following components:

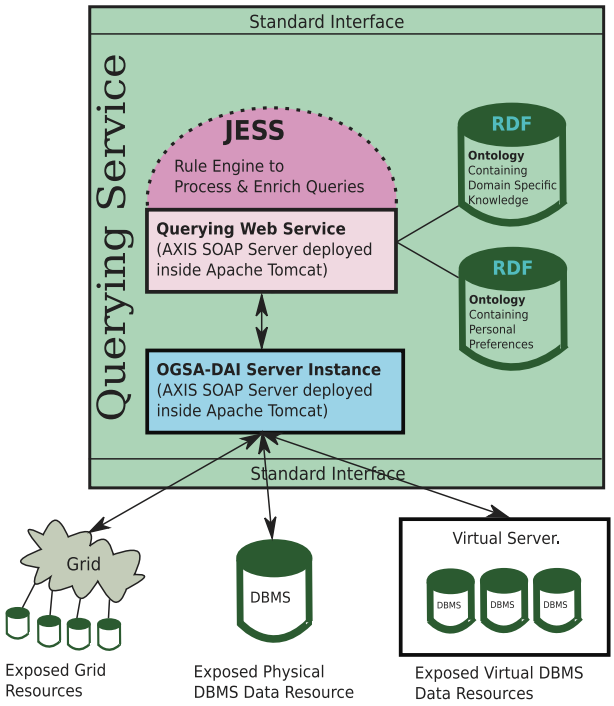

Figure 1: Querying service architecture.

- Rule-Engine -to process queries.
- Query Service -to interface with meta-data and data resources.
- Personal Meta-Data -providing user preference information.
- Domain Meta-Data -providing domain information.
- Heterogeneous, Distributed Data Resource Interface-to facilitate multi-modal data access/manipulation.

The rule-engine acts as the 'brain' of the service, it contains complex rules, which are configured according to data held as both personalization and domain-specific meta-data. The rules contained within this engine are used to dynamically interface the exposed resources through an OGSA-DAI server instance. The JESS Rule-Engine is used by the querying service to process and enrich queries. Rule chains are generated and stored inside a JESS instance using the RDF ontology containing domain-specific knowledge and personal preferences. Rule chains are a powerful way of representing our intelligence, this is because we can handle a wide variety of possible query scenarios without explicitly catering for each of them. The process of asserting facts and rules firing to generate actions can be observed in Figure 2. The querying service in neuGRID will take the form of a web service, which is compliant with the project's design philosophy [12]. This will act mainly as an interface providing access to the ontology encapsulating the personalization and domain-specific semantic representations. It is in this component that the API is exposed for external access. The personal meta-data

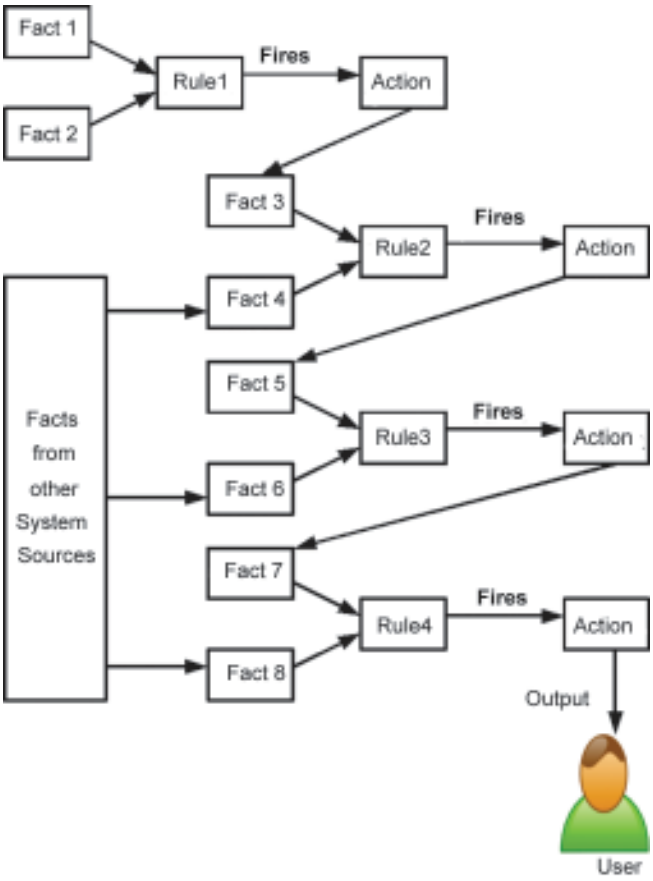

Figure 2: An example rule chain.

store contains user-specific information in the form of a user profile. This might include research interests in the context of a research tool, search settings and histories. Whilst the domain semantics are required to describe the domain, the user personalization semantics are required to give a user context within the system. When the personalization meta-data is combined with the meta-data concerning the domain, the querying service will tailor itself to the individual needs of the user.

5. Test-Bed Implementation

In order to evaluate the proposed querying architecture, a test-bed has been implemented in which technologies have been matched to the various components discussed previously. The following paragraphs describe the test-bed and briefly discuss the technologies used to produce the implementation. The aim of this is to demonstrate that the design is architecturally sound and show that all components communicate as anticipated with basic rules, which can be expanded. The architecture allows a range of clients to connect to the querying service using a standard interface provided by the SOAP web service. The test-bed features a browser-based client and an iPhone client. The querying interface is text based to keep the interface as simple as possible.

A Java web service has been developed to make up the querying service; this is set out in the neuGRID design philosophy, and the JESS [13] rule-engine was chosen since it is fast and tightly coupled with elements of the Java language. In order to represent the ontology containing the meta-data, RDF is used. The ontology can be edited using a tool such as protégé and parsed using the JENA API from within the querying service. The contents of the ontology denote the rules, which are created and implanted in the rule-engine to process queries as they are asserted as facts within the engine. Currently in existence in the service are rules handling the enrichment of queries through synonym awareness. This can be seen as the filtering of queries in order that they be correctly processed by the system e.g. A keyword search is processed as a keyword search and not as an SQL query. The advantage of using this approach containing dynamic ontology and a precompiled query service is that changes are made to ontology and not to the code of the service. This offers up the potential for real-time reconfigurability and therefore reduces the probability of system down time.

The storage layer provides access to all the resources that are exposed through the querying service. Most resources, which one would wish to connect to can be exposed through OGSA-DAI [14], but other abstraction layers may also be employed. Since the architecture is flexible, there are other options one could consider for this layer, another widely used possibility is AMGA [15]. For the test-bed, we chose OGSA-DAI since it represents a mature product with an impressive existing user base. Most of these heterogeneous data interfaces are implemented as a web service as is OGSA-DAI and it is invoked over SOAP from the querying service. This tool has a wealth of connectors to access various types of resources but it is open source thus additional methods can be added.

6. Results and Evaluation

Building on the deployed test-bed, some initial experimentation and results have been undertaken to make an assessment of the current implementation and architecture. This has involved the creation of a chain of fairly basic rules that handle several data types and enrich queries in a simple way. The service has been configured and deployed for the neuGRID environment and data resources. This was done using a virtualized deployment of the LORIS schema. LORIS is an e-infrastructure that is specific to medical imaging domain; it provides data management for images as well as storing clinical variable data for analysis in research studies. The schema represents the kind of data that will be used, containing information about clinical research projects and study subjects. In order to query such data, a concept has

been devised, which aims to process plain text from the simplest of 'Google-like' interfaces.

One of the first experiments was to implement some rules that harness domain-specific synonymic knowledge for basic query expansion. The synonymy knowledge is contained within an ontology and a set of rules are generated based on the content. These rules are used to pre-process the query before it is submitted to OGSA-DAI and the results are returned. Consider an example in which a researcher is interested in the pyriform lobe, which is a specific part of the brain. Terms are often synonymous in the biomedical field for cultural and historical reasons. A user does not want to waste time searching for all of the possibilities or risk missing out some important data. The query enhancement rules use domain-specific information to find four synonymous terms and to return the aggregated results for each to the user.

To provide query type detection, the ontology contains information about a particular type of query and how to recognize it. An SQL statement can be detected with the use of certain keywords and the syntax of the querying language. Once the correct type has been identified, the rule-engine determines how and where the query should be distributed and processed. This leads to the query being executed over the distributed test resources that have been configured and the results are returned to the user. Whilst these examples are somewhat simplistic, they do demonstrate that the core architecture and test-bed are functioning correctly. We now plan to develop more complex rules, which will inter weave increasing personalization with a greater level of domain-specific knowledge. These initial results, however, demonstrate a test-bed implementation of the architecture deployed with functionality, which can be built upon as required.

7. Future Work

The querying requirements of neuGRID would appear to have significant overlap with what may be necessary for querying the Future Internet. Data is likely to become more complex and be represented in varying ways, thus a flexible architecture is required to adapt to these dynamic data types. An example of data, which is likely to become increasingly prevalent in the future is live information feeds from sensor networks. This clearly presents a challenge for current querying mechanisms as the feeds are likely to be distributed and dynamic in nature. The initial results from the querying service may suggest that with further work, it could be configured to address some of these challenges. We aim to continue to explore how the querying service can become functionally rich, but in a transparent way to the user. The goal is to allow users to interact with the Future Internet in familiar ways by keeping the interface constant and as simple as possible even as data continues to evolve rapidly. The use of semantic meta-data, personalization and domain-specific information, coupled with a rule-based approach would appear to have the level of flexibility that is necessary to achieve this.

A possible and exciting way of expanding the system is to wrap the querying service within a multi-agent environment. Here agents would harness the knowledge and intelligence contained within the current service-based architecture to conduct queries on the behalf of users. This may pave the way for users to configure personalized individual agents that navigate the complexities of the Future Internet on their behalf. In doing this, chains of highly complex rules may need to be enriched by machine learning and other relevant techniques. A user could be defined within an ontology thus giving them context, and this tree of the ontology could be parsed by the web service to derive agent behavior. Further experimentation is key to the development of this architecture and the test-bed that has been created will greatly aid this process.

8. Conclusion

In this paper, an architecture is presented, which allows heterogeneous, distributed querying encompassing semantic enrichment and personalization mechanisms for query processing. There seems to be significant similarities between what is required in querying within the neurological domain and what may be necessary to query the emerging 'Internet of things' or Future Internet. Users do not wish to become familiar with highly complex querying languages and as such, we propose the use of rule-engines to determine how queries should be processed. A rule-engine is configured with rules from two sets of meta-data containing information about the domain and the user. The flexibility of the architecture is important if one is to embrace the growing number and type of resources in such systems. The generic nature of the architecture means that no commitments are made in terms of technology and this is important if the model is to be re-implemented in other domains. We have suggested that adding personalization could provide results of greater relevance especially in research scenarios. Through the implementation of a test-bed, we have demonstrated the powerful role that a rule-engine can play in the processing of queries. We have also discussed how the architecture could in the future be harnessed by agents, indeed this is one such direction the service could take.

9. Acknowledgments

The authors acknowledge the financial support of the Framework Program 7 of the EC through the Grant Agreement number

211714. In addition, they thank the partners in neuGRID for their contributions to this paper from Fatebenefratelli (Brescia, Italy), UWE (Bristol, UK), MaatG knowledge (Archamps, France), VUmc (Amsterdam, Netherlands), HealthGrid (Clermont, France), Prodema (Bronschhofen, Switzerland), CF consulting (Milan, Italy) and the Karolinska Institute (Stockholm, Sweden).

References

1. A. Nijholt. Socio-Technical Implementation: Socio-technical Systems in the Context of Ubiquitous Computing, Ambient Intelligence, Embodied Virtuality, and the Internet of Things. In: Handbook of Research on Socio-Technical Design and Social Networking Systems. IGI Global Books,Hershey, PA, USA, pp. 489-92.
2. M. J. Eppler, and J. Mengis. The Concept of Information Overload - A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines, The Information Society: An International Journal, 20(5), 2004, pp. 1-20.
3. H. Knublauch. Ontology-Driven Software Development in the Context of the Se-mantic Web: An Example Scenario with Protege/OWL, In Frankel, D.S., Kendall, E.F., McGuinness, D.L., eds.: 1st International Workshop on the Model-Driven Semantic Web (MDSW2004), 2004.
4. M. Maedche, and S. Staab. Ontology Learning for the Semantic Web, Intelligent Systems, IEEE, Vol. 16, Issue2, Mar.-Apr. 2001, pp. 72-9.
5. P. Bloodsworth, S. Greenwood, and J. Nealon. A Generic Model for Distributed Real-Time Scheduling Based on Dynamic Heterogeneous Data, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, Intelligent Agents and Multi-Agent Systems, vol. 2891/2003, pp.110-21.
6. M. Hammer, and D. McLeod. On Database Management System Architecture, Tech. Rep. MIT/LCS/TM-141, Massachusetts Institute of Technology, Cambridge MA, Oct. 1979.
7. A.P. Sheth, and J.A. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, ACM Computing Surveys, vol. 22, no. 3, Sep. 1990.
8. http://www.wolframalpha.com. Accessed April 28, 2009.
9. R.J. Bayardo *et al*., InfoSleuth: Agent-based semantic integration of information in open and dynamic environments, In proceedings of the1997ACMSIGMOD, May. 1997.
10. M.N. Huhns. From DPS to MAS to: Continuing the Trends,AAMAS2009Bu-dapest.
11. N.H. Shah, D.L. Rubin, and I. Espinosa, *et al*., Ontology-driven indexing of public datasets for translation bioinformatics, AMIA Summit on Tranlsational Bioinformatics, 2008, San Francisco.
12. neuGRID Design Philosophy http://www.neugrid.eu/. Accessed April 28, 2009.
13. JESS http://www.jessrules.com/. Accessed April 28, 2009.
14. OGSA-DAI http://www.ogsadai.org.uk/. Accessed April 28, 2009.
15. AMGA http://amga.web.cern.ch/amga/. Accessed April 25, 2009.

AUTHORS

Tom Lansdale just completed his Bachelors degree at the University of the West of England, Bristol and now works at CERN where he has a fellowship in the Fabric Development section in IT responsible for the Extremely Large Fabric Management Tool suites; in particular his work focuses on LEMON (LHC Era Monitoring). LEMON monitors every computer in the CERN Computer Center providing visualizations of metrics such as CPU load and managing alarm triggers.

E-mail: Thomas.Hector.Lansdale@cern.ch

Peter Bloodsworth is a Research Fellow in the Center for Complex Cooperative Systems at the University of the West of England, Bristol. His research is mainly within the field of Artificial Intelligence and in particular concerns the applications of semantics and multi-agent systems to provide personalized intelligent services. He is currently working on several European projects and has active collaborations with major research institutes including CERN, Great Ormond Street Hospital and other partners.

E-mail: peter.bloodsworth@cern.ch

Ashiq Anjum is working at University of the West of England and is doing his research work in collaboration with CERN on middleware development projects. He is also contributing to various European and American research projects and is an active researcher in the area of Grid and distributed systems. He has authored more than 30 peer-reviewed publications.

E-mail: Ashiq.Anjum@cern.ch

Irfan Habib is a research student in the Center for Complex Cooperative Systems at the University of the West of England. His work at present involves the scalable enactment of compute and data intensive workflows in Grid environments.

E-mail: Irfan.Habib@cern.ch

Yasir Mehmood is a research student in the Center for Complex Cooperative Systems at the University of the West of England. His research thus far focuses on provenance aware validation and modeling of distributed e-Science workflows.

E-mail: Yasir.Mehmood@gmail.com

Richard McClatchey is the Director of the Center for Complex Cooperative Systems and a CERN Associate, Professor McClatchey was notably a founding member of the center for eScience Research in Bristol. He is a Fellow of the Institute of Electrical Engineers (FIEE) and of the British Computer Society (FBCS). His specialties are in the management of distributed data and processes, Grid Computing and the design of data models for the capture of workflow information.

E-mail: Richard.Mcclatchey@cern.ch

DOI: 10.4103/0256-4602.55274; Paper No TR 211_09; Copyright © 2009 by the IETE