2009 Fifth International Conference on Semantics, Knowledge and Grid

Knowledge Engineering for PharmacoGenomic Molecular Imaging of the Brain

Carl Taswell

Global TeleGenetics, Inc.
Ladera Ranch, CA 92694, USA
ctaswell@computer.org

***Abstract*—Schizophrenia, Alzheimer's disease, Parkinson's disease, and other neuropsychiatric degenerative disorders and dementias impose an enormous economic and psychosocial burden on society, communities, and families. In order to gain a better understanding of gene-brain-behavior relationships, improve treatment, and find cures for these diseases, translational research must be conducted with clinical trials of new drugs and other interventions followed by genotyping and imaging biomarkers for patients with these neuropyschiatric degenerative disorders. This research, involving pharmacogenomic molecular imaging of the brain, will be extremely costly in many ways. Therefore, knowledge engineering with effective software tools and applications built upon a semantic-enabled informatics infrastructure remains a necessary prerequisite to facilitate a reduction of those research costs by maximizing the benefit obtained from existing data and minimizing the cost of generating new data. A knowledge engineering framework that serves this goal must operate in a cross-disciplinary manner that integrates data from diverse biomedical fields while at the same time incorporating the relevant computational mathematics, statistics, and informatics analyses for productive data mining.**

I. Introduction

Schizophrenia, Alzheimer's disease, Parkinson's disease, and other neuropsychiatric degenerative disorders and dementias impose an enormous economic and psychosocial burden on society, communities, and families. They exact a staggering toll in costs with some estimates reaching a trillion dollars by 2050 based on aging of the baby boomer generation in American society. Because Alzheimer's disease and mild cognitive impairment are the most prevalent of the older-adult-onset disorders, they have received much attention with extensive investigations including those studying genetic causes [1], molecular imaging [2], [3] and pharmacological treatment [4]. However, significant advances in understanding genotype-phenotype relationships have also been made recently for schizophrenia [5], [6] as one of the younger-adult-onset disorders.

Given current hypotheses on etiologies and the current absence of cures, early detection remains a critical component of any intervention strategy designed to delay or retard the decline (ie, shift the onset or slow the rate of decline) in cognition and function associated with these disorders. With the movement toward predictive, preventive, and personalized medicine, translational research with clinical trials that seek to leverage the combined power of genomics, proteomics, and metabolomics coupled with functional molecular imaging will yield important new insights and knowledge with which to discover disease-modifying agents and then develop and monitor treatment regimens. Advances in technologies for simultaneous PET-MRI [7] and molecular-genetic imaging based on reporter gene expression [8] as well as the recent successes of human brain imaging studies on the relationships between drugs, genotypes, molecular imaging markers, and behavioral phenotypes [5], [6], [9]–[13] bring new hope and the promise of a bright future for the nascent field of pharmacogenomic molecular imaging (PGMI) of the brain [14]. As a consequence of the complexity of gene-brain-behavior relationships and the accelerating generation and accumulation of massive amounts of multi-scale multi-modal data (see Fig. 1), an informatics infrastructure for the management and analysis of this data becomes an absolute necessity for progress in our understanding of the brain and efforts to intervene successfully in lessening the harm caused by neuropsychiatric disorders.

II. Informatics for Brain PGMI

The scientific study of gene-brain-behavior relationships with brain PGMI and the medical goal of improving the diagnosis and treatment of neuropsychiatric disorders together serve as the guiding motivation and medical scientific problem context for the development of the informatics infrastructure and software applications pursued by the author as described in prior work [14], [16]–[19]. Broadly stated, the current aims of this work in biomedical informatics can be summarized as the following projects:

1) To complete development and implementation of an interoperable informatics infrastructure for brain PGMI and the study of gene-brain-behavior relationships.
2) To demonstrate operational performance of the infrastructure, client-server tools, and a knowledge engineering workbench application for brain PGMI with exploratory scientific analysis (data mining) of gene-brain-behavior relationships using data from existing databases of rodent, primate, and/or human brain scans, gene expression activities, subject behavior characteristics, and/or other genotype-phenotype data associated with the image libraries.
3) To perform cost analyses and computer simulations of prospective clinical trials with brain PGMI for patients with neurodegenerative disorders in order to optimize the cost-effectiveness of clinical trial study designs.

978-0-7695-3810-5/09 $26.00 © 2009 IEEE
DOI 10.1109/SKG.2009.101

IEEE computer society

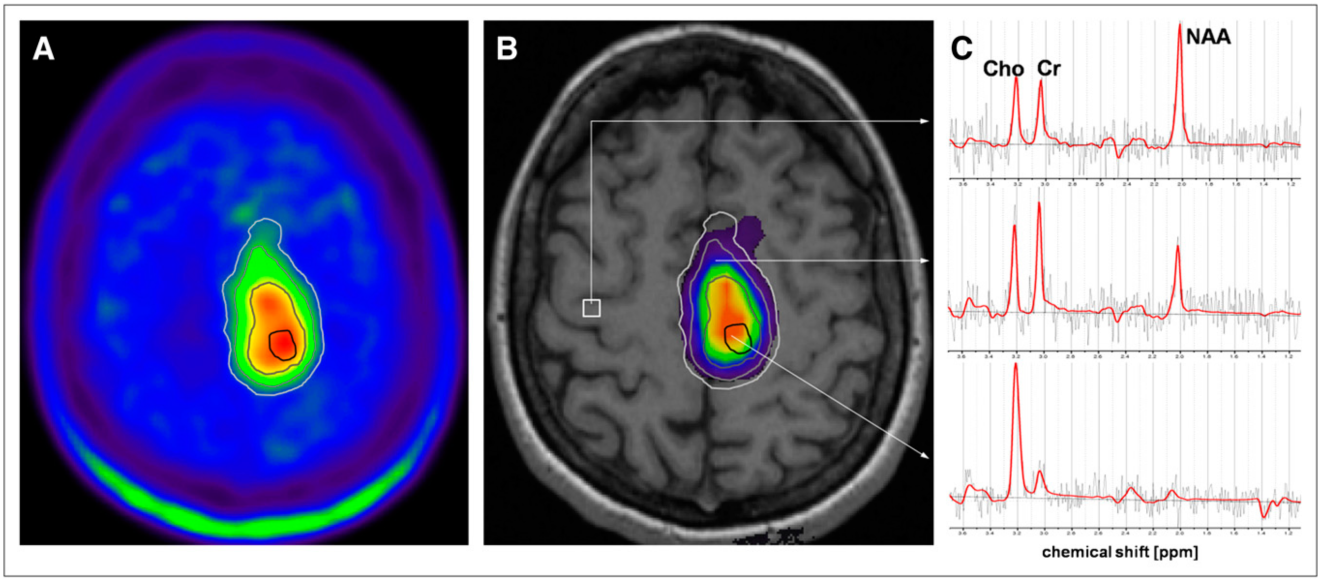


Fig. 1. Example of trimodal brain imaging with co-registration of a biological activity scan (PET with ^{18}F-FET radiotracer), structural anatomy scan (MRI with gadolinium contrast), and chemical spectroscopy scan (^{1}H MRSI); complete data set encompasses entire volume of the patient's brain; original figure published by Stadlbauer et al. [15, Fig. 2, p. 724] and reprinted by permission of the Society of Nuclear Medicine.

Prospective randomized controlled clinical trials of pharmacologic interventions followed by imaging biomarkers (including studies with new experimental radiopharmaceuticals) for patients with neuropyschiatric disorders who also undergo genotype analyses (including in vitro analysis of blood or sputum samples or pathological analysis of post-surgical or post-mortem brain tissue) could easily cost 10's, if not 100's, of millions of dollars especially if conducted as long-term multi-center international trials. Therefore, the preparatory work to build an informatics infrastructure, conduct data mining explorations, and perform cost analyses and computer simulations remains a *necessary prerequisite* not only to derive maximal information and benefit from existing data but also to reduce the costs of producing new data and gaining a better understanding of gene-brain-behavior relationships. Only then will actual clinical trials and the research enterprise be more productive in serving the public health need to provide better care for patients with neuropsychiatric disorders.

III. Current Neuroinformatics Systems

The book *Neuroinformatics* edited by Crasto [20] provides a compendium of informatics for brain science and medicine covering neuroscience knowledge management, computational neuronal modeling, brain imaging, and applications in neurogenetics for neurodegenerative disorders. Progress has been made by privately funded projects such as the brain-map.org portal for brain gene expression activity mapping of the Allen Institute for Brain Science as well as by publicly funded projects such as the neuinfo.org portal for the Neuroscience Information Framework of the NIH Blueprint for Neuroscience. Despite continuing efforts to address the existing barriers to interoperability for current data stores and to begin the transition to a semantic web of meaningfully linked and integrated data [21], the challenging task of reaching the goal of truly interoperable data has only just begun and many hurdles remain in the way.

In fact, current neuroinformatics portals still remain essentially isolated from other portals because there is no uniform standardized shared interface for all of the portals in neuroinformatics and related biomedical sciences to communicate with each other. Thus, it remains necessary to interact individually and separately with each portal's custom interface. Moreover, general initiatives (i.e., not specific to neuroinformatics) such as ncbcs.org/biositemaps and linkeddata.org represent short-term fixes that may help temporarily with some of the problems but will not suffice as a long-term solution for all of the problems (see below in Section IV). Most importantly, *none* of the current neuroinformatics portals or applications have yet been designed to focus on the use of pharmacogenomic molecular imaging for clinical trials. Finally, as reflected by the compendium *Neuroinformatics* [20], there remains a divide between the neuroinformatics for management of data from scientific experiments and the neuroinformatics for pseudo-realistic modeling of neurons and neuronal networks without a concerted effort to bridge this gap. However, there does exist a common mathematical model of network graphs that can characterize both the neural pathways of a living brain and the messaging pathways of the PORTAL-DOORS System [17] as a core informatics infrastructure. Studying the similarities and differences between the living brain network and the engineered communications network should enable a better understanding of both.

IV. A New Approach

The author has pursued a new approach distinguished by its goal of building a distributed shared infrastructure rather than a single centralized site. The design for the infrastructure core, called the PORTAL-DOORS System for the semantic web [17], was modeled on the enormously successful design of the IRIS-DNS System for the original web. More specifically, the Internet Registry Information Service (IRIS) registers domain names while the Domain Name System (DNS) publishes domain addresses with mapping of names to addresses for the original web. Analogously, the Problem Oriented Registry of Tags And Labels (PORTAL) registers resource labels and tags while the Domain Ontology Oriented Resource System (DOORS) publishes resource locations and descriptions with mapping of labels to locations for the semantic web. Both the IRIS-DNS System and the PORTAL-DOORS System share a common architectural style for pervasive metadata networks that operate as distributed metadata management systems with hierarchical authorities for entity registering and attribute publishing. Hierarchical control of metadata redistribution throughout the registry-directory networks constitutes an essential characteristic of this architectural style called Hierarchically Distributed Mobile Metadata (HDMM) with its focus on moving the metadata for *who what where* as fast as possible from servers in response to requests from clients [22].

PORTAL-DOORS and IRIS-DNS each operate as information-seeking support systems that function as hierarchical registry-directory networks for the distribution of mobile metadata. While the original motivation for the design of PORTAL-DOORS has been and remains that of serving the goals of neuroinformatics for the study of gene-brain-behavior relationships, PORTAL-DOORS was also designed to solve several major problems of web engineering: cybersilos in scientific discourse, search engine consolidation, registry/repository centralization, and barriers to progress in the transition from original web to semantic web [19].

It should be noted that the original design of the internet protocols and IRIS-DNS systems were configured purposefully to promote failsafe redundancy and high-speed efficiency for distributed communication networks, and that there are significant risks when consolidation or centralization result in monopolistic control of centralized hubs as single points of failure. Moreover, non-hierarchical peer-to-peer strategies may work fine for delivery of large amounts of data to a known destination, but they will not work for search and discovery of a small datum at an unknown destination within a large universe of data. Thus, it should also be noted that simple flat non-hierarchical or peer-to-peer approaches (such as the web site files of Biositemaps or the data and web page markup of LinkedData) will not scale to meet the demands presented by the ever accelerating growth in production of data, web pages, and web sites and services. Only a properly structured hierarchical approach akin to the successful design of IRIS-DNS will scale sufficiently to meet the demands of the explosive growth in data required to solve the problems of brain diseases and disorders. This data will be available in varying kinds whether public or private, raw or processed, analyzed as qualitative or quantitative results, interpreted as inferred conclusions, redacted for publication in literature, etc., and thus, amenable to data mining and/or text mining to varying degrees.

To use a geographic metaphor, simple non-hierarchical approaches risk trapping the information seeker stuck and bogged down in the valleys of isolated lands around a world in which any information sought and found in an isolated valley is not shared with or redistributed to other isolated valleys. In contrast, properly designed and structured hierarchical approaches enable the information seeker to send message requests efficiently from any valley to the closest mountain peak and then from that mountain peak to other mountain peaks surveying all valleys in all lands (see Figure 2) in order to efficiently obtain the requested data which then automatically becomes shared and redistributed in other valleys and lands as part of the response to the request. Such redistribution and sharing of information does not occur in the non-hierarchical approaches of initiatives like Biositemaps and LinkedData.

To use another metaphor, the simple non-hierarchical approaches lack the ability to scale and solve the worsening problems of finding needles in haystacks which can only be solved by the more sophisticated, versatile, and flexible hierarchical approaches of systems that implement the architectural style common to both IRIS-DNS and PORTAL-DOORS. Therefore, the PORTAL-DOORS System, as the core infrastructure system for the biomedical informatics work pursued by the author, maintains the same principles of architectural design that have been so successfully tested and proven by IRIS-DNS for decades. In this regard, PORTAL-DOORS represents a dramatically different approach from all other current initiatives (which are all non-hierarchical) whether intended for the semantic web in general or for neuroinformatics portals in particular.

V. Architectural Design of PORTAL-DOORS

In accordance with the HDMM architectural style, PORTAL-DOORS has been designed to serve the semantic web and grid in a manner analogous to the way that IRIS-DNS has served the original web. The design from the original 'blueprint' paper [17] has been updated with revisions [18], [19]. Note that the original *separate* design of PORTAL registries and DOORS directories has been supplemented with a new bootstrapping *combined* design with integrated NEXUS registrars [19]. Both can coexist together.

Table I summarizes some of the similarities and differences between the IRIS-DNS and PORTAL-DOORS paradigms from the perspective of considering both as distributed database systems with entity registering and attribute publishing implemented with the HDMM architectural style [22]. Technical details of the PORTAL-DOORS paradigm are further elaborated in the publications [17]–[19], [22] and at portaldoors.org. Some important characteristics include:

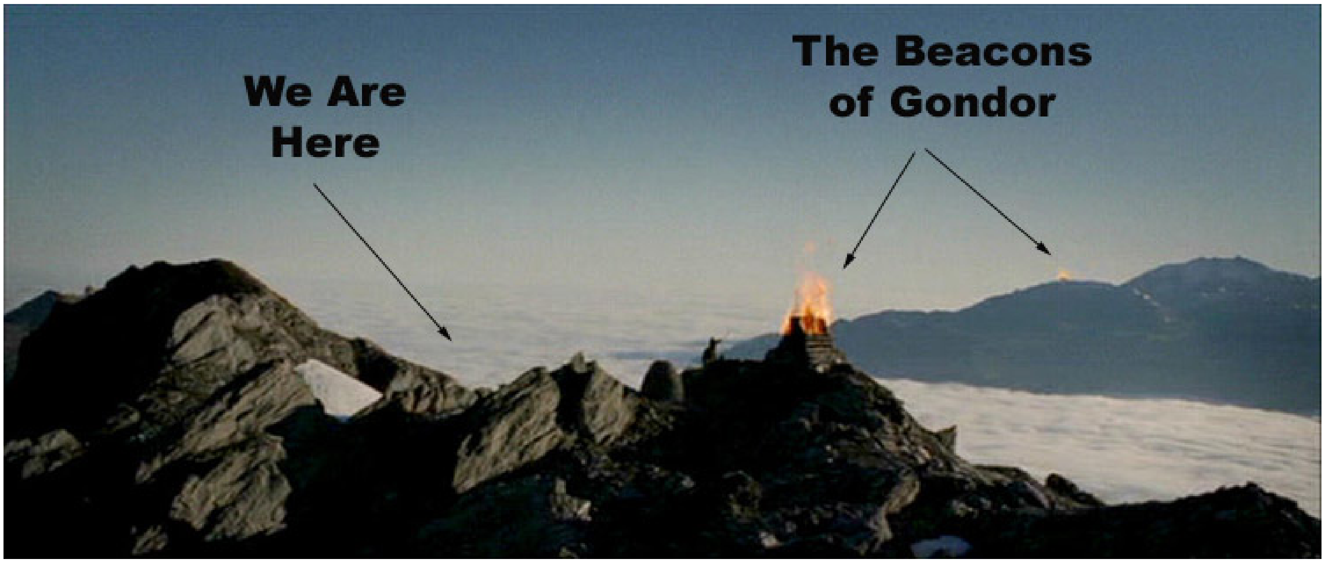


Fig. 2. Beacons of Gondor dramatize a metaphor for the advantages of hierarchical communication networks that enable search and discovery of a small item in a very large world. If everybody remains trapped under the clouds in isolated valleys everywhere and unable to see elsewhere, then how will we (or software agents) communicate with each other fast enough to find and reach unknown destinations, persons (or agents), and small pieces of information in a large world that grows ever larger all the time?

TABLE I
HIERARCHICALLY DISTRIBUTED MOBILE METADATA SYSTEMS WITH ENTITY REGISTERING AND ATTRIBUTE PUBLISHING

	IRIS-DNS System	PORTAL-DOORS System
Dynamic metaphor	A distributed communications network brain of nodal neurons continuously updating, exchanging, and integrating messages about ‘who what where’	
Static metaphor	A simple phonebook	A sophisticated library card catalogue
Registering system	IRIS registries	PORTAL registries
— Entity registered	domain	resource
— Identified by	unique name	unique label (URI or IRI) with optional tags
Publishing system	DNS directories	DOORS directories
— Attributes published	address and aliases	location and descriptions
— Specified by	IP number	URIs, URLs, RDF triples referencing OWL ontologies
Forwards requests	Yes	Yes
Caches responses	Yes	Yes
Serves original web	Yes via mapping of character name to numeric address	Yes via mapping of character label to URL for IRIS-DNS
Serves semantic web	No (IRIS-DNS does not use RDF triples)	Yes via mapping of character label to semantic description
Crosslinks entities	No	Yes via mappings within DOORS descriptions to other resources
Crosslinks systems	No	Yes via mappings within PORTAL crossreferences to other systems

- A distributed network of registries and directories for resource metadata oriented by problem domain or specialist community rather than by technology format of the resource.
- A hierarchical system enabling local independence of communities while simultaneously maintaining global compatibility for communication between and search amongst different communities.
- A hybridized architecture with both XML Schemas and terminologies serving the original web and also RDF triples and OWL ontologies serving the semantic web to bridge and transition from the original web to the semantic web.
- Decentralization, distribution, and democratization to promote evolutionary adoption of componentized terminologies and ontologies (i.e., survival of the fittest, not necessarily the first).
- Hierarchical authorities and globally unique identifiers to prevent namespace conflicts when identifying resources while maintaining autonomy of local communities with control over local policies.
- Designed to accomodate any resource — whether abstract or concrete, offline or online, semantic or non-semantic — with either non-semantic descriptions using tags refer-

encing terminologies or semantic descriptions using RDF triples referencing ontologies.

- Supported with cross-references to other systems whether legacy or contemporaneous.

The PORTAL-DOORS System is *not* another attempt once again to create a so-called "one stop shop" that claims to be the "one and only" destination for "all shopping needs". In fact, the general philosophy of HDMM systems turns that notion upside down and argues that centralized "one stop shops" cannot and will not solve the problems. Instead, there should be a multiplicity and diversity of registries and directories (as well as other kinds of sites and portals) continuously exchanging mobile metadata. The requirement for mobility of this metadata mandates that the metadata become highly distributed, redistributed, and cached everywhere for the speed and efficiency of search and location which can be achieved effectively only by maintaining the interoperability of all registries and directories to communicate with each other transparently within the same infrastructure system.

VI. Infrastructure vs. Tools vs. Content

PORTAL-DOORS as a lower-level infrastructure system must be distinguished from higher-level tools and applications built on the foundation of the infrastructure. PORTAL-DOORS as a mobile metadata management, communication, and distribution system must also be distinguished from the actual metadata as the content that the infrastructure is designed to send, receive, and exchange throughout the system. Fundamentally, the PORTAL-DOORS System establishes an interoperable, platform-independent, application-independent, interface standard for information exchange over the internet with a design that is guided by the HDMM architectural style, specified to fulfill additional requirements to serve both the original web and semantic web as described in the design 'blueprint' paper [17], and currently partially detailed in a draft reference implementation written in XML Schema *.xsd files.

Work to complete a reference implementation must clarify not only the structural data model for metadata records, but also the functional behavioral model for the PORTAL and DOORS services in response to requests from clients. Servers and clients must also communicate over transport protocols. The PORTAL-DOORS Project maintains a vision of serving more than one transport protocol as discussed in Section VII.E. of [17]. Initial drafts (prior to version 0.5) assumed use of the IRIS core protocol. The current draft (version 0.5) addresses only the structural data model. The next draft (version 0.6) will re-introduce use of a specific transport protocol but replace the IRIS core protocol with an http protocol using RESTful web services. At present, in a bootstrapping stage of development for PORTAL-DOORS, RESTful web services do provide a more favorable environment for promoting adoption of the system. However, a fully dedicated and optimized protocol specifically for PORTAL-DOORS may ultimately prove necessary to achieve the speed and efficiency comparable to that which exists now for IRIS-DNS.

As PORTAL-DOORS continues to be developed and implemented, any tool, application, or web site that accesses PORTAL-DOORS must be distinguished from the system itself. The PORTAL-DOORS System should not be considered either a single site or repository any more than the IRIS-DNS System of domain name registries and directories could be construed to be a single site or repository. For both IRIS-DNS and PORTAL-DOORS infrastructure systems, server data stores and client tools and applications can be written in any language on any platform. Client tools are necessary for agents to edit the information maintained at an individual server data store. Client tools are also necessary for agents and users to navigate, search and query the information stored not only at a particular server but also throughout the entire network of servers. These tools include faceted browsers, keyword search utilities, and SPARQL query interfaces.

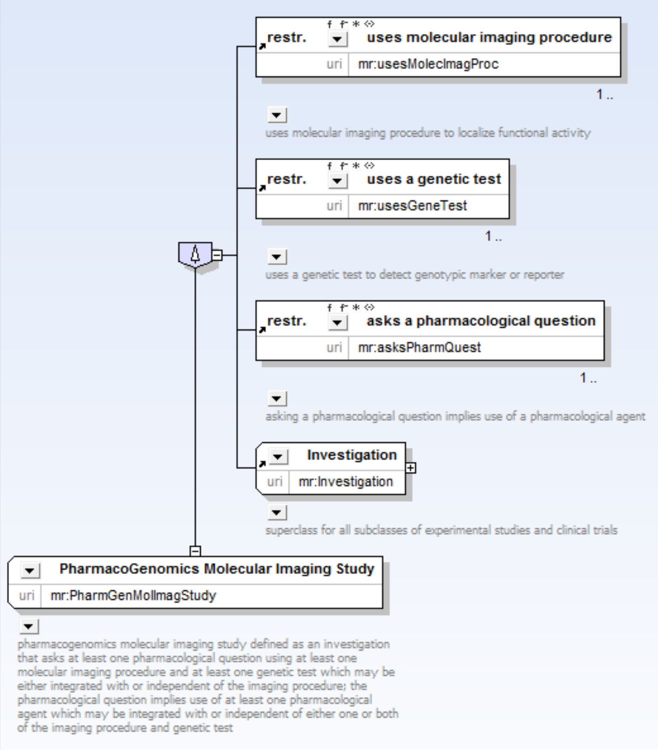
Even more complex applications can be built in which the navigation, search, and query tools may be embedded within more sophisticated applications that hide these tools from the user interface. An important example is an application component that would provide natural language answers to natural language questions in the context of the overall function of the software application. In this example, the component converts the user's natural language question to a SPARQL query submitted to PORTAL-DOORS, and then converts the query response from PORTAL-DOORS back to a natural language answer for presentation to the user.

The usefulness of any technology system designed to manage content, regardless of how it is constructed from interface standards, server networks, client tools, and applications, is only as good as the content that it manages and exposes to producers and consumers of the content. Without this content exposed by the system, the system itself remains of limited practical utility. Thus, generation of content remains an important aspect of the development of any content management system. At present with a web browser interface, entry of metadata records into PORTAL-DOORS is performed by human agents much akin to the manner of entry for metadata records into IRIS-DNS.

However, software agents such as webbots and converters could be developed which would be able to generate metadata records for resources automatically. Presumably, there would be a trade-off in the quality of content produced versus the rate of content production when comparing records created automatically by software agents with records curated by human agents. This trade-off would not be applicable to those situations where an existing structured database only needs an appropriate interface for inbound queries and wrappers for outbound responses in order to expose metadata records for resources contained within the database.

VII. Workbench for Brain PGMI

Successful design and development of a knowledge engineering workbench for brain PGMI would produce a critical enabling software application for informatics research relevant to brain PGMI and the study of gene-brain-behavior



uses molecular imaging procedure to localize functional activity

uses a genetic test to detect genotypic marker or reporter

asking a pharmacological question implies use of a pharmacological agent

superclass for all subclasses of experimental studies and clinical trials

pharmacogenomics molecular imaging study defined as an investigation that asks at least one pharmacological question using at least one molecular imaging procedure and at least one genetic test which may be either integrated with or independent of the imaging procedure; the pharmacological question implies use of at least one pharmacological agent which may be integrated with or independent of either one or both of the imaging procedure and genetic test

Fig. 3. Formal class definition for a pharmacogenomic molecular imaging study excerpted from the ManRay ontology implemented in OWL.

relationships (see Sec. II). However, it should be built upon the foundation of the PORTAL-DOORS infrastructure system of networked registries and directories in order to maintain interoperability according to the vision of the PORTAL-DOORS paradigm. This paradigm favors a flexible and modular approach promoting collaborative networks of cross-linking resources and inter-referencing ontologies in a multiplicity of problem oriented domains that may or may not be conceptually related. Moreover, it is necessary to build a minimal set of those registries and directories that would facilitate knowledge engineering for the cross-disciplinary field of brain PGMI.

Prototype registries have been developed within PORTAL-DOORS that are specialized for various problem-oriented domains relevant to brain PGMI: the GeneScene registry for genetics, ManRay for nuclear medicine, BrainWatch for brain imaging and neuropsychiatry, and BioPORT for biomedical computing. These registries facilitate translational biomedical informatics for brain PGMI by assuring globally unique identification of resources while promoting interoperability and enabling cross-registry searches between the different specialty fields that contribute to pharmacogenomics, molecular imaging, and brain imaging.

BioPORT demonstrates a simple example where the current purpose remains limited to publishing the availability of biomedical computing resources (see [17] for definition and scope). ManRay demonstrates a more complex example with the multiple goals of cataloguing the components, defining the protocols, and interlinking the patient registries that will enable clinical trials with pharmacogenomics molecular imaging studies (see [16], [23]). In particular, since PORTAL registries require constraints defining the problem-oriented domain, the ManRay registry restricts registration of resource entities to those that can be lexically and/or semantically related to nuclear medicine and molecular imaging. It permits these entities to be anything from a person or organization involved with the field to imaging agents and devices used in the field including radiopharmaceuticals, protocols and scanners. Further, the ManRay ontology associated with the ManRay registry contains the formal OWL class definition of a generic PGMI study (see Fig. 3). Ontology class definitions such as this generic example enable derivative constructs for more specific examples when combined with restrictions using classes from other ontologies such as the BrainWatch ontology associated with the BrainWatch registry, necessary for the class definition of a brain PGMI study.

When fully developed and implemented, a knowledge engineering workbench for brain PGMI should be able to drive from question to answer through a semantic web of linked data guided by the metadata map of interconnecting PORTAL-DOORS registries and directories. This workbench should also serve as a knowledge management system for imaging radiotracers and biomarkers that enable mapping the neural pathways of the brain and elucidating genotype-phenotype correlations relevant to neuropsychiatric disorders. Thus, it could be extended with customized registries for use in multi-center clinical trials tracking patients and the associated multi-scalar multi-modal molecular imaging libraries from gene expression array images to brain scan images. Different registries could be maintained for each of the necessary study components whether qualified investigational sites, the radiopharmaceutical INDs and CMCs, the patients, or the images themselves. In all cases, the investigator would determine whether a type of entity is registered as a resource according to the focus of interest and requirements of the situation. Moreover, even if a particular registry is maintained privately for confidentiality, it could still benefit from interaction with other registries that have been exposed publicly. Therefore, a knowledge engineering workbench could provide an effective environment for use with public and private data, metadata, and information.

VIII. Provenance and Reproducibility

A knowledge engineering workbench for brain PGMI must be able to search the metadata maps of the PORTAL-DOORS networks and then to access and analyze data obtained from the resources identified and described in the metadata maps. To do so, the workbench should provide, utilize, and integrate the diverse kinds of data with a comprehensive collection of computing components that address all major aspects of computational imaging by incorporating the necessary mathematics, statistics, and informatics including both numeric- and semantic-based artificial intelligence methods. Three fundamental classes of data processing can be considered for an image: preprocessing algorithms for image generation, processing algorithms for image management, and postprocessing algorithms for image analysis. Images from data libraries may be published as the raw data requiring reconstruction prior to visualization, processed data available as conventional images, or analyzed data available as extracted and quantitated feature sets. All of the image data (whether from brain scans, gene expression arrays, or other sources of images) must be further integrated, analyzed, and correlated with the non-image data included within the study.

A virtual workbench such as the one just described should enable an investigator to conduct scientific research according to the classical tenets of the experimental method. Concerns about these matters have led to the recent popularity of discussions and papers on workflow [24], integration [25], provenance [26]–[28], and reproducibility [29], [30]. However, in the context of knowledge engineering and semantic data integration, ontology alignment remains a significant challenge [31] for which progress is hindered by overuse of the same term or phrase for different concepts. Such overuse can even be considered misleading when it results in confusion about the meaning of terms across disciplines and the confounding of issues in science and the conduct of interdisciplinary scientific research. In this regard, both the term "provenance" and phrase "reproducible research" have become problematic.

Use of the term provenance has become so pervasive in some authors' work that almost everything seems to have become a form of provenance. One is left wondering whatever happened to terms such as materials, methods, protocols, procedures, functions, algorithms, programs, etc., and why such terms should be neglected or abandoned. If concerned about the challenges of ontology matching and alignment, knowledge engineers should also remain concerned about maintaining distinctions in use of terms so that different terms are appropriately used for different concepts. This common sense practical usage should occur both more formally in ontologies and less formally in speech and writing.

In this regard, there is an important benefit in not confounding the term "provenance" with a multiplicity of uses but instead reserving it solely for tracking the chain of custody of an object from owner to owner or other custodians, curators, and users of the object. Doing so does not confound that use of the term with either the creation of the object or with the subsequent manipulation of the object by an individual for which there is no relevant concept of transfer of custody of the object. Here creation and/or manipulation of the object, whether by an algorithm or within a workflow of successive manipulations by multiple componentized algorithms, are all controlled by the same investigator. One only need to reflect upon the art world where the term originated to appreciate that an artist would never use the term "provenance" to refer to the process by which he created his own work of art.

A number of authors have also been using the term "reproducibility" and the phrase "reproducible research" in a manner that risks confounding important principles in science. Their usage of these words appear to reduce to advocacy for several practices: 1) clear and explicit communication in a scientific paper about materials and methods, 2) distribution of software as a part of the methods, and 3) distribution of data sets as part of the materials. While certainly laudable practices to be encouraged, they are not new. Empirical statisticians as a community (distinct from theoretical statisticians) have a long tradition of publishing data sets and then testing and comparing various statistical methods on those data sets. Numerical analysts as a community (now also known as computational mathematicians or computational scientists) have a long tradition of clearly and explicitly communicating the mathematical equations, numerical methods, algorithms, pseudo-code, and actual software code along with detailed error analyses and performance metrics for best-case, worst-case, average-case scenarios, etc.

Thus, when authors such as [29], [30] use the term "reproducibility" and the phrase "reproducible research," they are actually addressing questions such as the following: How clearly and explicitly did the research paper communicate materials and methods? Is an electronic copy of the materials and methods available in an executable format that reproduces another electronic copy of the research paper? But these questions are distinct from the critical question of whether the reported scientific conclusion is reproducible by an *independent* investigator using *independent* materials and methods or an *independent* approach. Moreover, re-executing the same digital code that reproduces the same digital result only proves that a machine re-computed the same output from the same input. It does not answer any questions about whether the calculated results are correct or incorrect, or the inferred conclusions valid or invalid. Therefore, alternative concepts and terms must be distinguished and maintained separately as defined formally by [32], [33] for *repetitive executability*, *input-output repeatability*, and *scientific reproducibility* in the context of inputs (parameters and raw data) and outputs (results and/or processed data) for computational algorithms.

More generally, the terms "reproducible" or "reproducibility" should not be used in a way that detracts from the single most important meaning of and question about reproducibility in science: Is the scientific conclusion reported by the research paper true? For example, even if a paper claims to be "reproducible" in the sense of [29], [30] but demonstrates a result for only a single test case or a few so-called "toy examples", then it does not follow as a necessary consequence that the scientific conclusion reported is true and reproducible over an entire population or for other classes of similar data (see [34] for further discussion). Thus, claims about reproducibility of papers, methods, results, and conclusions must always be interpreted with respect to the definition assumed for the term reproducibility. The logical alternative appropriate for knowledge engineering and semantic data integration is simply to use distinct terms for distinct concepts. Doing so helps solve semantic data integration problems.

IX. Conclusion

The current status and future plans of the PORTAL-DOORS Project have been reviewed in this report with respect to the HDMM architectural style, the PORTAL-DOORS System itself, and the important use case of knowledge engineering for brain PGMI and the study of gene-brain-behavior relationships. Interdisciplinary science knowledge engineering issues relevant to system construction (infrastructure vs. tools vs. content) and to semantic data integration (provenance and reproducibility) have also been reviewed and discussed.

References

[1] N. Ertekin-Taner, "Genetics of Alzheimer's disease: A centennial review," *Neurol Clin*, vol. 25, no. 3, pp. 611–667, 2007.

[2] L. Mosconi, W. H. Tsui *et al.*, "Multicenter standardized F18-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease, and other dementias," *J Nuc Med*, vol. 49, no. 3, pp. 390–398, mar 2008.

[3] N. Tolboom, M. Yaqub, W. M. van der Flier, R. Boellaard, G. Luurtsema, A. D. Windhorst, F. Barkhof, P. Scheltens, A. A. Lammertsma, and B. N. M. van Berckel, "Detection of Alzheimer pathology in vivo using both 11C-PIB and 18F-FDDNP PET." *J Nucl Med*, vol. 50, no. 2, pp. 191–197, Feb 2009. [Online]. Available: http://dx.doi.org/10.2967/jnumed.108.056499

[4] K. M. Sink, K. F. Holden, and K. Yaffe, "Pharmacological treatment of neuropsychiatric symptoms of dementia: A review of the evidence," *JAMA*, vol. 293, no. 5, pp. 596–608, 2005.

[5] S. J. Huffaker, J. Chen, K. K. Nicodemus, F. Sambataro, F. Yang, V. Mattay, B. K. Lipska, T. M. Hyde, J. Song, D. Rujescu, I. Giegling, K. Mayilyan, M. J. Proust, A. Soghoyan, G. Caforio, J. H. Callicott, A. Bertolino, A. Meyer-Lindenberg, J. Chang, Y. Ji, M. F. Egan, T. E. Goldberg, J. E. Kleinman, B. Lu, and D. R. Weinberger, "A primate-specific, brain isoform of KCNH2 affects cortical physiology, cognition, neuronal repolarization and risk of schizophrenia." *Nat Med*, vol. 15, no. 5, pp. 509–518, May 2009. [Online]. Available: http://dx.doi.org/10.1038/nm.1962

[6] K. L. Narr, P. R. Szeszko, T. Lencz, R. P. Woods, L. S. Hamilton, O. Phillips, D. Robinson, K. E. Burdick, P. Derosse, R. Kucherlapati, P. M. Thompson, A. W. Toga, A. K. Malhotra, and R. M. Bilder, "DTNBP1 is associated with imaging phenotypes in schizophrenia." *Hum Brain Mapp*, May 2009. [Online]. Available: http://dx.doi.org/10.1002/hbm.20806

[7] M. S. Judenhofer, H. F. Wehrl, D. F. Newport, C. Catana, S. B. Siegel, M. Becker, A. Thielscher, M. Kneilling, M. P. Lichy, M. Eichner, K. Klingel, G. Reischl, S. Widmaier, M. Rcken, R. E. Nutt, H.-J. Machulla, K. Uludag, S. R. Cherry, C. D. Claussen, and B. J. Pichler, "Simultaneous PET-MRI: a new approach for functional and morphological imaging." *Nat Med*, vol. 14, no. 4, pp. 459–465, Apr 2008. [Online]. Available: http://dx.doi.org/10.1038/nm1700

[8] J. H. Kang and J.-K. Chung, "Molecular-genetic imaging based on reporter gene expression." *J Nucl Med*, vol. 49 Suppl 2, pp. 164S–179S, Jun 2008. [Online]. Available: http://dx.doi.org/10.2967/jnumed.107.045955

[9] C. Ernst, A. Sequeira, T. Klempan, N. Ernst, J. Ffrench-Mullen, and G. Turecki, "Confirmation of region-specific patterns of gene expression in the human brain." *Neurogenetics*, vol. 8, no. 3, pp. 219–224, Aug 2007. [Online]. Available: http://dx.doi.org/10.1007/s10048-007-0084-2

[10] A. V. Witte, A. Flel, P. Stein, M. Savli, L.-K. Mien, W. Wadsak, C. Spindelegger, U. Moser, M. Fink, A. Hahn, M. Mitterhauser, K. Kletter, S. Kasper, and R. Lanzenberger, "Aggression is related to frontal serotonin-1a receptor distribution as revealed by PET in healthy subjects." *Hum Brain Mapp*, Dec 2009. [Online]. Available: http://dx.doi.org/10.1002/hbm.20687

[11] B. Draganski, S. A. Schneider, M. Fiorio, S. Klppel, M. Gambarin, M. Tinazzi, J. Ashburner, K. P. Bhatia, and R. S. J. Frackowiak, "Genotype-phenotype interactions in primary dystonias revealed by differential changes in brain structure." *Neuroimage*, vol. 47, no. 4, pp. 1141–1147, Oct 2009. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2009.03.057

[12] A. Lothe, C. Boni, N. Costes, P. Gorwood, S. Bouvard, D. L. Bars, F. Lavenne, and P. Ryvlin, "Association between triallelic polymorphism of the serotonin transporter and ^{18}F-MPPF binding potential at 5-HT(1A) receptors in healthy subjects." *Neuroimage*, May 2009. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2009.04.067

[13] E. M. van de Giessen, M. M. L. de Win, M. W. T. Tanck, W. van den Brink, F. Baas, and J. Booij, "Striatal dopamine transporter availability associated with polymorphisms in the dopamine transporter gene SLC6A3." *J Nucl Med*, vol. 50, no. 1, pp. 45–52, Jan 2009. [Online]. Available: http://dx.doi.org/10.2967/jnumed.108.053652

[14] C. Taswell, "PORTAL-DOORS infrastructure system for translational biomedical informatics on the semantic web and grid," in *Proceedings of the American Medical Informatics Association Summit on Translational Bioinformatics, San Francisco, CA*, Mar 2008, p. 43.

[15] A. Stadlbauer, O. Prante, C. Nimsky, E. Salomonowitz, M. Buchfelder, T. Kuwert, R. Linke, and O. Ganslandt, "Metabolic imaging of cerebral gliomas: spatial correlation of changes in O-(2-^{18}F-fluoroethyl)-L-tyrosine PET and proton magnetic resonance spectroscopic imaging." *J Nucl Med*, vol. 49, no. 5, pp. 721–729, May 2008. [Online]. Available: http://dx.doi.org/10.2967/jnumed.107.049213

[16] C. Taswell, B. Franc, and R. Hawkins, "The ManRay project: Initial development of a web-enabled ontology for nuclear medicine," in *Proceedings of the 53rd Annual Meeting of the Society of Nuclear Medicine, San Diego, CA*, Jun 2006, p. 1431.

[17] C. Taswell, "DOORS to the semantic web and grid with a PORTAL for biomedical computing," *IEEE Trans Inform Technol Biomed*, vol. 12, no. 2, pp. 191–204, Mar 2008, in the "Special Issue on Bio-Grid".

[18] ——, "Implementation of prototype biomedical registries for PORTAL-DOORS," in *Proceedings of the American Medical Informatics Association Summit on Translational Bioinformatics, San Francisco, CA*, Mar 2009, AMIA-0036-T2009.

[19] ——, "Alternative bootstrapping design for the PORTAL-DOORS cyberinfrastructure with self-referencing and self-describing features," in *Semantic Web*. Vienna, Austria: IN-TECH Publishing, 2009, in press.

[20] C. J. Crasto, Ed., *Neuroinformatics*, ser. Methods in Molecular Biology. Humana Press, 2007, vol. 401.

[21] T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, and D. J. Weitzner, "Creating a science of the web." *Science*, vol. 313, no. 5788, pp. 769–771, Aug 2006.

[22] C. Taswell, "The hierarchically distributed mobile metadata (HDMM) style of architecture for pervasive metadata networks," 2009, submitted.

[23] ——, "Application of the PORTAL-DOORS system for use by clinical trials registries," in *Proceedings of the American Medical Informatics Association Summit on Translational Bioinformatics, San Francisco, CA*, Mar 2009.

[24] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, "Examining the challenges of scientific workflows," *Computer*, vol. 40, no. 12, pp. 24–32, Dec 2007.

[25] P. A. Bernstein and L. M. Haas, "Information integration in the enterprise," *Commun. ACM*, vol. 51, no. 9, pp. 72–79, 2008.

[26] L. Moreau, P. Groth, S. Miles, J. Vazquez-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan, and L. Varga, "The provenance of electronic data," *Commun. ACM*, vol. 51, no. 4, pp. 52–58, 2008.

[27] J. Freire, D. Koop, E. Santos, and C. T. Silva, "Provenance for computational tasks: A survey," *Computing in Science and Engineering*, vol. 10, no. 3, pp. 11–21, 2008.

[28] A. J. Mackenzie-Graham, J. D. V. Horn, R. P. Woods, K. L. Crawford, and A. W. Toga, "Provenance in neuroimaging." *Neuroimage*, vol. 42, no. 1, pp. 178–195, Aug 2008. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2008.04.186

[29] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden, "Reproducible research in computational harmonic analysis," *Computing in Science and Engineering*, vol. 11, no. 1, pp. 8–18, 2009.

[30] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *Signal Processing Magazine, IEEE*, vol. 26, no. 3, pp. 37–47, May 2009.

[31] J. Euzenat and P. Shvaiko, *Ontology matching*. Heidelberg (DE): Springer-Verlag, 2007. [Online]. Available: www.ontologymatching.org

[32] C. Taswell, "Specifications and standards for reproducibility of wavelet transforms," in *Proceedings of the International Conference on Signal Processing Applications and Technology*. Miller Freeman, Oct. 1996, pp. 1923–1927.

[33] ——, "Reproducibility standards for wavelet transform algorithms," Computational Toolsmiths, Tech. Rep. CT-1998-01, Mar. 1998. [Online]. Available: www.toolsmiths.com/docs/CT199801.pdf

[34] ——, "Experiments in wavelet shrinkage denoising," *Journal of Computational Methods in Sciences and Engineering*, vol. 1, no. 2s–3s, pp. 315–326, 2001.