NPDSLINKS: Nexus-PORTAL-DOORS-Scribe Learning Intelligence aNd Knowledge System

1st Shreya Choksi
Brain Health Alliance
Ladera Ranch, CA, USA
schoksi@bhavi.us

2nd Peter Hong
Brain Health Alliance
Ladera Ranch, CA, USA
phong@bhavi.us

3rd Sohyb Mashkoor
Brain Health Alliance
Ladera Ranch, CA, USA
smashkoor@bhavi.us

4th Carl Taswell
Brain Health Alliance
Ladera Ranch, CA, USA
ctaswell@bhavi.us

***Abstract*—With the continuing growth in use of large complex data sets for artificial intelligence applications (AIA), unbiased methods should be established for assuring the validity and reliability of both input data and output results. Advancing such standards will help to reduce problems described with the aphorism 'Garbage In, Garbage Out' (GIGO). This concern remains especially important for AIA tools that execute within the environment of interoperable systems which share, exchange, convert, and/or interchange data and metadata such as the *Nexus-PORTAL-DOORS-Scribe* (NPDS) cyberinfrastructure and its associated *Learning Intelligence aNd Knowledge System* (LINKS) applications. The PORTAL-DOORS Project (PDP) has developed the NPDS cyberinfrastructure with lexical PORTAL registries, semantic DOORS directories, hybrid Nexus diristries, and Scribe registrars. As a self-referencing and self-describing system, the NPDS cyberinfrastructure has been designed to operate as a pervasive distributed network of data repositories compliant with the Hierarchically Distributed Mobile Metadata (HDMM) architectural style. Building on the foundation of the NPDS cyberinfrastructure with its focus on data, PDP has now introduced LINKS applications with their focus on algorithms and analysis of the data. In addition, PDP has launched a pair of new websites at NPDSLINKS.net and NPDSLINKS.org which will serve respectively as the root of the NPDS cyberinfrastructure and the home for definitions and standards on quality descriptors and quantitative measures to evaluate the data contained within NPDS records. Prototypes of these descriptors and measures for use with NPDS and LINKS are introduced in this report. PDP envisions building better AIA and preventing the unwanted phenomenon of GIGO by using the combination of metrics to detect and reduce bias from data, the NPDS cyberinfrastructure for the data, and LINKS applications for the algorithms.**

***Index Terms*—Semantic web, knowledge engineering, data stewardship, metadata management, quality metrics, PORTAL-DOORS Project, NPDS cyberinfrastructure, LINKS applications.**

I. Introduction

Charles Babbage, inventor of the first calculating machines, described his interactions with others when he presented his *difference engine* to the members of England's Parliament in the early 19th century [1], [2]:

> "On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' … I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."

The simple intuitive principle implied by that description has remained central to the core foundation of calculating and computing machines from the early history of primitive computers to the present era with the advances of multi-core chip architectures, big data, and artificial intelligence.

Over a century after Babbage made his famous remarks, Army Specialist William Mellin expressed his concern about the inability of computers to think for themselves when interviewed for a 10 November 1957 newspaper article, and explained that "sloppily programmed" inputs inevitably lead to incorrect outputs [3]. The Hammond Times newspaper of Hammond Indiana published Mellin's explanation of this concept of flawed data producing flawed results with the phrase "Garbage In, Garbage Out" and the acronym GIGO.

Even with a theoretically perfect computer model or computing machine, absence of quality in the input data yields a consequential absence of quality in the output results (see Figure 1 on GIGO), where the term quality here serves as shorthand for the phrase validity and reliability. Thus, it remains necessary to develop and maintain standards for reviewing and curating the quality of data before using and applying the data when asking and answering research questions involving that data, and when evaluating the functionality and operations of a cyberinfrastructure system with a network of computing nodes and data repositories.

In this report, we discuss the current implementation of the *Nexus-PORTAL-DOORS-Scribe* (NPDS) cyberinfrastructure, introduce our associated *Learning Intelligence aNd Knowledge System* (LINKS) applications built on the NPDS foundation, and propose initial versions of anti-GIGO descriptors and measures for the records stored in the NPDS data repositories which will be analyzed by the algorithms of the LINKS applications. These descriptors and measures have been defined to evaluate not only individual fields separately for each of the lexical PORTAL and semantic DOORS components of NPDS, but also collectively the status of all PORTAL fields, all DOORS fields, and all Nexus fields representing the entire infoset for a resource entity. Implementation of a diversity of descriptors and measures characterizing the quality and quantity of data in NPDS repositories will support greater confidence in appropriate inferences made about results obtained from LINKS applications with artificial intelligence, machine learning, and/or expert systems that analyze the data.

DOI: 10.1109/TransAI2020???

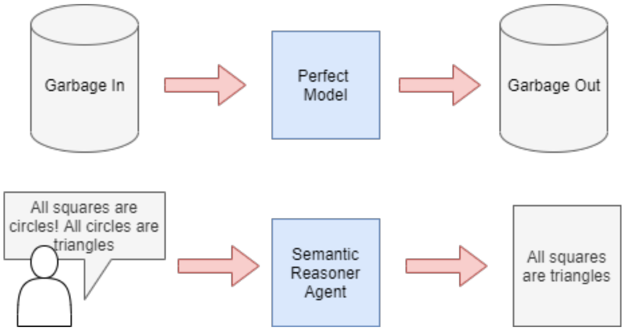
Garbage In
Perfect Model
Garbage Out
All squares are circles! All circles are triangles
Semantic Reasoner Agent
All squares are triangles

Figure 1. Garbage In, Garbage Out.

II. NPDS Cyberinfrastructure with Data

As scientific data accumulates larger in size, more complex in scope, and widespread in distribution accompanying the development of more powerful computers and computing technologies, the human enterprise of scientific research will require the assistance and support of AIA to search, consume, parse, and analyze this data. Approaches to the design and development of algorithms for analysis of data with AIA vary both in their use of quantitative and qualitative methods, and also in their use of mathematical, statistical, logical and/or ontological tools. AIA built with the XML, RDF, and OWL technology stack of the semantic web continue to confront challenges during the ongoing transition from the lexical web. Some of these transition barriers include the inaccessibility of data in isolated data silos and slow adoption of common message exchange standards instead of a distributed open communicating network of interoperable data repositories [4]. Together with concerns about the consolidation of search engines into an effective oligopoly (perhaps even a de facto monopoly?) and the spread of misinformation and malinformation [5], these obstacles have limited the growth of the semantic web and constrained the distribution of information in a desired knowledge network necessary to answer questions in various problem domains.

The PORTAL-DOORS Project (PDP) designed the original PORTAL-DOORS cyberinfrastructure for registering resource entities and publishing attributes about them, as a distributed system modeled in analogy with the IRIS-DNS system [4]. Originally proposed by Taswell in 2006, the Problem Oriented Registry of Tags and Labels (PORTAL) built for the lexical web serves to register resource labels and tags analogous to IRIS registering domain names, and the Domain Ontology Oriented Resource System (DOORS) built for the semantic web serves to publish resource locations and descriptions analogous to DNS publishing numerical addresses corresponding to the domain names [4]. Since its origin in 2006, PDP has been pursued to develop the Nexus-PORTAL-DOORS-Scribe (NPDS) cyberinfrastructure, which serves as a "who what where" diristry-registry-directory system for identifying, describing, locating, and linking things on the internet, web, and grid. Based on the Hierarchically Distributed Mobile Metadata (HDMM) style of architecture for pervasive metadata networks [6], NPDS serves the original vision of resource entity data and metadata publishing, albeit enhanced from the original separation of concerns with lexical PORTAL registries and semantic DOORS directories now to the hybrid Nexus diristries [5] and combined Scribe registrars [7]. NPDS offers a distributed and decentralized infrastructure system by allowing individuals and organizations to maintain independent repositories of semantic and lexical metadata with data for and about resource entities in different problem domains of interest.

The design principles for PDP and NPDS [4], [5] have been renamed the DREAM principles [8] where the acronym DREAM represents the comprehensive summarizing phrase "Discoverable Data with Reproducible Results for Equivalent Entities with Accessible Attributes and Manageable Metadata". Within this collection of concepts realized in the PDP-DREAM ontology, the phrase "Equivalent Entities" as a shortened version of the question "Equal or Equivalent Entities?" [9] represents the principle of paramount importance to the conduct of scientific research as the essential enquiry of identifying and characterizing two entities as either the same, similar, related, or different from each other [10]. Moreover, this principle remains applicable not only to entities in experimental scientific research such as hypotheses, data, results, inferences, and claims in the published literature, but also to practical management of replicate or separate records in database management systems. When should two records be preserved separately because they represent different entities and when should they be merged because they represent redundant representations of the same entity? It will be important to distinguish between true equivalence and false equivalence when evaluating the sameness versus similarity of entities prior to considering whether to merge their database records.

III. LINKS Applications with Algorithms

We refer to the Nexus-PORTAL-DOORS-Scribe cyberinfrastructure of distributed network repositories of data as the *NPDS cyberinfrastructure with data*. Analogously, we refer to the associated Learning Intelligence aNd Knowledge System applications for analysis of the data as the *LINKS applications with algorithms*. Thus, we use the acronym AIA for artificial intelligence applications in general, while we use the acronym LINKS for those AIA developed specifically by PDP for NPDS. Both terms, *NPDS* and *LINKS*, may be prefixed with *PDP-* as *PDP-NPDS* and *PDP-LINKS*, and may also be combined together with each other as in the title of this report with the term *NPDSLINKS*. Moreover, we have launched the web site at www.NPDSLINKS.net to serve as the root of the NPDS cyberinfrastructure with a Scribe registrar intended for NPDS components. Recall that *components* are defined in the NPDS nomenclature as entities representing the network servers including the Nexus diristries, PORTAL directories, DOORS directories, and Scribe registrars, whereas *constituents* are defined as entities representing persons or organizations who are the agents, owners, and/or registrants of the entities [5]. An accompanying website at www.NPDSLINKS.org will serve as home to our PDP work on LINKS applications with algorithms

DOI: 10.1109/TransAI2020???

including the development of quality descriptors and quantity measures to evaluate the NPDS cyberinfrastructure with data (see Section IV).

We envision that the desired synergy between the NPDS cyberinfrastructure with data and the LINKS applications with algorithms will generate a productive knowledge engineering system. Knowledge systems have been described by Alavi *et al.* as environments "developed to support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application" [11]. The interaction of LINKS algorithms with NPDS data, as an effective learning intelligence and knowledge system, will exploit the exchange of NPDS messages through its distributed network of registries, directories, and diristries, thereby facilitating storage, retrieval, and transfer of knowledge as NPDS records between and across different problem-oriented domains. Thus, NPDS and LINKS will continue to be developed to provide a synergistic open system for information search and retrieval by which investigators can readily explore transdisciplinary resources which are not restricted to a single problem-oriented domain.

With NPDSLINKS and continuing efforts by PDP to encompass a wide variety of problem-oriented domains in the biomedical sciences, different research communities should be able to communicate with each other and learn from each other. Neuroscience, one of the original application domains for development of the NPDS cyberinfrastructure, serves as an example of a field that can benefit from interaction with the related field of machine learning. Artificial neural networks show great promise to generate models of brain function and behavior [12] because they remain analogous to networks of neurons that comprise the brain. Patterns of neural activity produced by artificial systems may reveal important insights on the brain's own functionality [13]. Taswell commented "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", thus studying the similarities and differences between the two could enable a better understanding [14]. With LINKS applications and cross-referencing resource entities interlinking within a distributed system of NPDS data repositories each designed for a specific field of research inquiry and managed by that research community, transdisciplinary bridges can be built in place of silo walls, and researchers can more readily examine and compare similarities and differences both within and across scientific fields.

IV. Descriptors and Measures of Data

All AIA, including the LINKS applications for NPDS data discussed in this report, remain critically dependent on the use of input data of sufficient quality and quantity to generate output results of possibly comparable quality and quantity. Certainly, the input data are not the sole determinant of the validity and reliability of the output results. Rather, input data are a necessary but not sufficient determinant of the validity and reliability of output results. Figure 1 demonstrates the consequences of false outputs for false inputs with the example of a semantic reasoning algorithm which analyzes the statements "all squares are circles" and "all circles are triangles". Obviously, both statements are false. However, if the semantic reasoning analyzer is not aware that the input statements are false and instead assumes that they are true, then it could deduce the output statement that "all squares are triangles". Thus, outputs can be as untrue as inputs. Similarly, in the domain of machine learning with neural networks, input layer nodes must process valid data in order for the neural network to learn and generate valid results from the output layer nodes. With the new big-data driven implementations of machine learning systems, Gudivada *et al.* declared that the greatest challenge for solving big-data problems remains the nature of the data itself and that "high-quality datasets are essential for developing machine learning models" [15]. Indeed, both the quantity and quality of both data and metadata remain essential in ensuring valid and reliable analyses and interpretation of the data with results from AIA.

To evaluate both quantitatively and qualitatively the content of records in the NPDS cyberinfrastructure data repositories, we propose both quantitative measures and qualitative descriptors not only for the individual fields, but also for the respective groups of fields, from each kind of NPDS record (*i.e.*, either Nexus, PORTAL, DOORS, or Scribe records). Logical indicators can be reported simply as true or false with boolean values. They reflect whether the concept or content tested is present or absent [16] as used in experimental design and data analysis to address the important problems of missing data and null or NaN values. Quantitative measures can be reported as simple counts with integer values of defined items in fields (declared as *required, permitted, or optional* [4], [5]), or as more sophisticated metrics with float values such as the FAIR family of ratio-based metrics for plagiarism detection [17]. They evaluate the nature and characteristics of the data beyond the simple question of presence versus absence. Qualitative descriptors can be reported as categorical variables with enum values. They can check the content for a level of compliance with a declared standard involving a specified regex, syntax, or serialization format such as XML, RDF, OWL, HTML, or XHTML. Such categorical variables may have ranks, scores, or values corresponding to the recommendations of the particular serialization format, or more generally, may be reported simply as one of the three values *none, lax, or strict* with respect to the compliance of the content to the standard. Moreover, the defined list of permitted terms for a categorical descriptor may also be declared by the administrators of the diristry for a particular problem-oriented domain, thus supporting flexibility and extensibility of analysis by specific research communities independently managing and curating their data repositories.

The explanation provided above for descriptors and measures of the data has been generic in the sense that it pertains to all fields of NPDS records. However, remarks concerning individual named fields serve to provide clarifying examples of evaluations specific to one or some but not all fields. For the DOORS Locations field, validation checks can assess different kinds of locations including both physical and

DOI: 10.1109/TransAI2020???

virtual addresses. Locations that are URL addresses can be resolved, pinged, and assessed for response media type as application, image, text, etc, and as JSON, XML, HTML, etc. Locations that are postal service mail addresses and geophysical addresses can be validated by a geolocation service for verification of the normalized mail delivery address as well as the latitude and longitude coordinates. For the DOORS Provenances field, validation checks can involve verifying cited sources and origins for the resource entity as in the example of bibliographies with cited references for those entities that represent publications in the scientific literature. For the DOORS Descriptions field, evaluations may involve a reasoning agent or engine that tests for non-contradictory logical consistency of claims in the content and also inferences for entailments from those content claims. Note that this reasoning analysis extends beyond the simpler validation checks for compliance with a syntax standard. As these examples demonstrate, descriptors and measures of the data unique to each of the named fields can be applied individually in addition to those that can be applied in a common generic manner to either all or groups of the NPDS record fields.

V. Conclusion

Throughout the history of computers and computing, data scientists and computational engineers have been aware of the problems associated with the GIGO phenomenon of “Garbage In, Garbage Out” [1]–[3]. In this report, we outlined some of our plans with the PDP for adopting anti-GIGO approaches when maintaining the integrity of the *NPDS cyberinfrastructure with data* and supporting the validity and reliability of the *LINKS applications with algorithms*. We described a variety of qualitative descriptors and quantitative measures, including logical indicators, simple counts, more sophisticated metrics, and categorical scores or ranks all of which can be used for evaluation of the nature of the content in our learning intelligence and knowledge system, both with respect to quantity and quality of the data. These anti-GIGO approaches will remain essential for our LINKS applications such as the examples described by Taswell *et al.* [18] with automated meta-analyses of the clinical trial literature. To follow our ongoing progress with the PDP on the NPDS cyberinfrastructure and LINKS applications, visit our new websites respectively at www.NPDSLINKS.net and www.NPDSLINKS.org.

References

[1] C. Babbage, *Passages from the Life of a Philosopher*. Good Press, 2019.

[2] R. Stenson, *Is this the first time anyone printed, ’garbage in, garbage out’?* Mar. 2016.

[3] W. Mellin, “Work with new electronic brains opens field for army math experts,” *The Hammond Times*, vol. 10, p. 66, 1957.

[4] C. Taswell, “DOORS to the semantic web and grid with a PORTAL for biomedical computing,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 2, pp. 191–204, 2 Mar. 2008, In the Special Section on Bio-Grid published online 3 Aug. 2007, ISSN: 1089-7771. DOI: 10.1109/TITB.2007.905861.

[5] ——, “A distributed infrastructure for metadata about metadata: The HDMM architectural style and PORTAL-DOORS system,” *Future Internet*, vol. 2, no. 2, pp. 156–189, 2010, In Special Issue on Metadata and Markup., ISSN: 1999-5903. DOI: 10.3390/FI2020156. [Online]. Available: www.mdpi.com/1999-5903/2/2/156/.

[6] ——, “The hierarchically distributed mobile metadata (HDMM) style of architecture for pervasive metadata networks,” in *2009 IEEE 10th International Symposium on Pervasive Systems, Algorithms, and Networks*, IEEE, Dec. 2009, pp. 315–320. DOI: 10.1109/I-SPAN.2009.73.

[7] A. Craig, S.-H. Bae, and C. Taswell, “Bridging the semantic and lexical webs: Concept-validating and hypothesis-exploring ontologies for the Nexus-PORTAL-DOORS System,” *Journal of Systemics, Cybernetics and Informatics*, vol. 15, no. 5, pp. 8–13, Jul. 11, 2017. [Online]. Available: www.iiisci.org/journal/sci/FullText.asp?id=BA947YN17.

[8] A. Craig, A. Ambati, S. Dutta, *et al.*, “DREAM principles and FAIR metrics from the PORTAL-DOORS Project for the semantic web,” in *2019 IEEE 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, (Jun. 28, 2019), Pitesti, Romania: IEEE, Jun. 2019. DOI: 10.1109/ECAI46879.2019.9042003. [Online]. Available: www.portaldoors.org/pub/docs/ECAI2019DREAMFAIR0612.pdf.

[9] A. Athreya, S. K. Taswell, S. Mashkoor, *et al.*, “Essential question: ‘equal or equivalent entities?’ about two things as same, similar, or different,” in *2020 IEEE 2nd International Conference on Transdisciplinary Artificial Intelligence (TransAI)*, (Sep. 21, 2020).

[10] S. Dutta, K. Uhegbu, S. Nori, *et al.*, “DREAM Principles from the PORTAL-DOORS Project and NPDS Cyberinfrastructure,” in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, IEEE, Feb. 4, 2020, pp. 211–216. DOI: 10.1109/ICSC.2020.00044. [Online]. Available: www.portaldoors.org/pub/docs/ECAI2019DREAMFAIR0612.pdf.

[11] M. Alavi and D. E. Leidner, “Knowledge management and knowledge management systems: Conceptual foundations and research issues,” *MIS quarterly*, pp. 107–136, 2001.

[12] M.-A. T. Vu, T. Adal, D. Ba, *et al.*, “A shared vision for machine learning in neuroscience,” *The Journal of Neuroscience*, vol. 38, no. 7, pp. 1601–1607, Jan. 2018.

[13] N. Savage, “How AI and neuroscience drive each other forwards,” *Nature*, vol. 571, no. 7766, S15–S17, Jul. 2019.

[14] C. Taswell, “Knowledge engineering for PharmacoGenomic Molecular Imaging of the brain,” in *2009 Fifth International Conference on Semantics, Knowledge and Grid*, Institute of Electrical and Electronics Engineers (IEEE), Sep. 2009, pp. 26–33. DOI: 10.1109/SKG.2009.101.

[15] V. Gudivada, A. Apon, and J. Ding, “Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations,” *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1–20, 2017.

[16] E. Babbie, *The practice of social research*. Belmont, CA: Wadsworth Cengage Learning, 2013, ISBN: 1-133-04979-6.

[17] A. Craig, A. Ambati, S. Dutta, *et al.*, “Definitions, formulas, and simulated examples for plagiarism detection with FAIR metrics,” in *2019 ASIS&T 82nd Annual Meeting*, (Oct. 19, 2019), vol. 56, Melbourne, Australia: Wiley, 2019, pp. 51–57. DOI: 10.1002/PRA2.6. [Online]. Available: www.portaldoors.org/pub/docs/ASIST2019FairMetrics0611.pdf.

[18] S. K. Taswell, A. Craig, D. Leung, *et al.*, “Hypothesis-exploring methods for automated meta-analyses of brain imaging literature,” in *Proceedings Annual Meeting of the Western Region Society of Nuclear Medicine*, Monterey CA, 2015. [Online]. Available: www.portaldoors.org/pub/docs/WRSNM2015TnT1p1020.pdf.

DOI: 10.1109/TransAI2020???