BRAINIACS
BRAIN IMAGING AND COMPUTING SCIENCES

2020 December
Volume 1 Issue 1 Edoc B61CA3D89

The Nexus-PORTAL-DOORS-Scribe (NPDS) Learning Intelligence aNd Knowledge System (LINKS)*

Shreya Choksi and Carl Taswell†

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 described 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.

Keywords

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

Contents

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 [2], [3]:

> "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 [4]. The Hammond Times newspaper of Hammond Indiana published Mellin's explanation of

*Document received 2020-Dec-07, published 2020-Dec-30. A preliminary version of this work was presented at the IEEE 2020 TransAI Conference [1].

†All authors are affiliated with Brain Health Alliance Virtual Institute, Ladera Ranch, CA 92694 USA; correspondence to CTaswell at Brain Health Alliance.

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 algorithm and 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 descriptor sets separately for each of the lexical PORTAL and semantic DOORS components of NPDS, but also collectively the status of all PORTAL descriptor sets, all DOORS descriptor sets, and all Nexus descriptor sets 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, or expert systems that analyze the data.

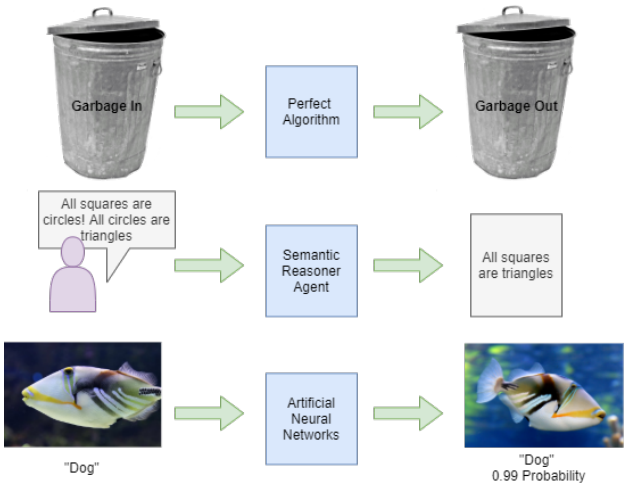

Figure 1: Garbage In, Garbage Out.

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 [5]. 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 [6], 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 [5]. 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 [5]. 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 [7], 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 [6] and combined Scribe registrars [8]. 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. These repositories range in problem-domain, several of which are brain health oriented diristries ranging in application from BrainWatch for brain imaging and neuropsychiatry [9] to HELPME for health, education, law, public policy, and medical ethics.

The design principles for PDP and NPDS [5], [6] have been renamed the DREAM principles [10] 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?" [11] 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 [12]. 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. Tools and systems for analyzing equivalence in both semantic and lexical terms will be a part of the AIA included in the LINKS system, which has important implications in both reproducibility and plagiarism.

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*. These applications include automated search and meta-analysis of published scientific literature where smart agents search and find resources for problem domains and then extract inferences from semantic metadata associated with those researches [13], as well as software agents with converters to automatically populate Nexus diristries [14]. 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 (see Figure 2). 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 [6]. An accompanying website at www.NPDSLINKS.org will serve as home to our PDP work on LINKS applications with algorithms including the development of quality descriptors and quantity measures to evaluate the NPDS cyberinfrastructure with data (see Section).

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" [15]. These knowledge management systems aim at "fostering learning processes, knowledge sharing, collaboration between knowledge workers irrespective of their location, etc" [16]. 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 [17] 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 [18]. 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 [9]. 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.

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. In the case of faulty training sets with mislabeled data, the Hawaiian reef triggerfish might be wrongly identified as a dog arising from confusion regarding the Hebrew word "Dag" which resembles the English "dog" but means "fish". Machine learning algorithms will then subsequently identify all future cases of the Hawaiian reef triggerfish as dogs when the two are meant to be distinct (see Figure 1). 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" [19]. Indeed, 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 individually named metadata property sets, but also for the respective

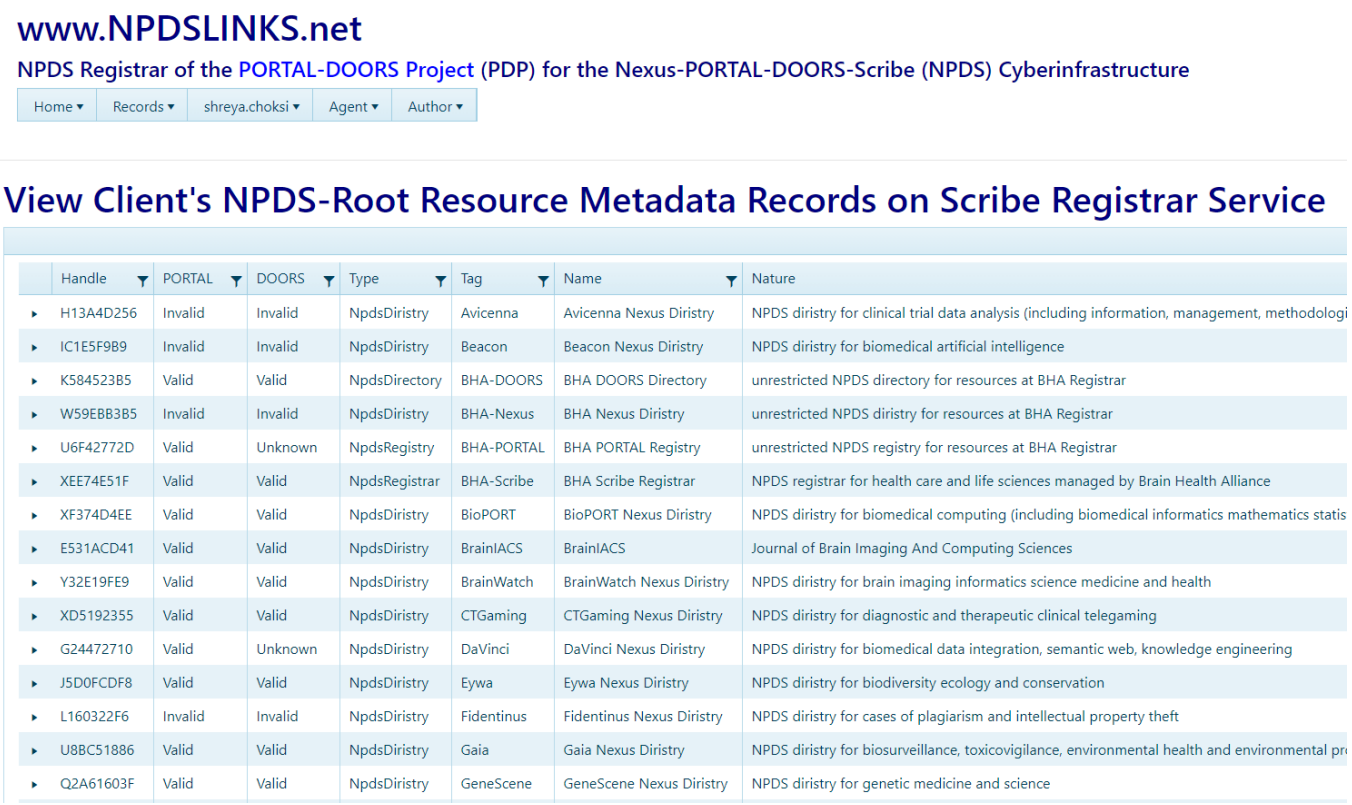

Figure 2: NPDS-Root at www.NPDSLINKS.net

group named metadata property sets from each kind of NPDS record (*i.e.*, either Nexus, PORTAL, DOORS, or Scribe records, see Tables 1 and 2). Quality measures will be comprised of a combination of indicators, quantitative descriptors, and qualitative descriptors. 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 [20] 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 the named metadata property sets (declared as *required, permitted, or optional* [5], [6]), or as more sophisticated metrics with float values such as the FAIR family of ratio-based metrics for plagiarism detection [21]. 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. As outlined by general use case scenarios in [6], usage of the different named metadata properties may range from minimal use of both required and permitted elements for each of PORTAL and DOORS to maximal use of both permitted and required elements for both PORTAL and DOORS combined as Nexus, as well as optional elements that may be declared by the repository administrator.

Figure 3 depicts an example of a quality record from the HELPME diristry with all PORTAL and DOORS metadata property sets from "EntityLabels" to "Distributions" containing some content. The Status tab displays the implementation of several of these quality measures. The number of items in a named property set has been expressed as a simple integer count for quantitative measures, while overall quality for each of PORTAL and DOORS infosets can be found expressed as "Infoset PORTAL Status" and "Infoset DOORS Status". In the case of the quality record in Figure 3, both Infoset Statuses are Valid. As a counterexample, Figure 4 demonstrates a poorly curated record with next to none of the named property sets containing any content, and the Status tab reflects this poor quality with item counts of zero and lack of validity on the PORTAL and DOORS Infoset Status.

PORTAL Metadata Quality Checks

The explanation provided above for descriptors and measures of the data has been generic in the sense that it pertains to all metadata properties of NPDS records. However, remarks concerning individual named metadata properties serve to provide clarifying examples of evaluations specific to one or some but not all metadata properties. Figure 5 shows a PORTAL metadata quality record from the HELPME diristry with all PORTAL metadata property sets containing some content. The status of these metadata property sets is displayed on the

Table 1: Quality Check Definitions

Symbol	Definition
I_P	Binary indicator for presence/absence of an element (or item) in a metadata property set (or list).
I_C	Binary indicator for whether the item is concept valid for a particular diristry.
I_R	Binary indicator for whether the item can be parsed using a defined regex.
Q_E	Simple integer count for the number of elements in a metadata property set.
Q_T	Integer counts for the number of RDF triples and inferences extracted from them.
C_S	Categorical descriptor for the syntax/serialization format of the element in the metadata property set.
C_L	Categorical descriptor for the latitude and longitude coordinates of a physical address using a geolocation service.
C_W	Categorical descriptor for the defined list of permitted element tag values specified in an enumerator.

Table 2: Quality Checks for NPDS Metadata Property Sets

Property Set	Indicator			Quantitative Descriptor		Categorical Descriptor		
	I_P	I_C	I_R	Q_E	Q_T	C_S	C_L	C_W
EntityLabels	Yes	No	Yes	Yes	No	No	No	No
SupportingTags	Yes	Yes	No	Yes	No	No	No	No
SupportingLabels	Yes	No	Yes	Yes	No	No	No	No
CrossReferences	Yes	No	Yes	Yes	No	No	No	No
OtherTexts	Yes	No	No	Yes	No	Yes	No	Yes
Locations	Yes	No	Yes	Yes	No	No	Yes	No
Descriptions	Yes	No	No	Yes	Yes	Yes	No	No
Provenances	Yes	No	No	Yes	No	No	No	Yes
Distributions	Yes	No	No	Yes	No	No	No	Yes

Status tab where each property set has a value for its "Status" when quality checked. We have adopted the term "Check" for those quality measures concerning the metadata property sets within each of the PORTAL and DOORS Infosets, while the term "Validate" is reserved for general groups of NPDS metadata property sets such as the PORTAL, DOORS, and Nexus Infosets.

For the PORTAL Infoset, the EntityLabels property set includes both the CanonicalLabel and AliasLabels which identify the resource entity, while the property sets for SupportingTags (as free text) and SupportingLabels (as URLs) provide supporting information for the resource entity in either the NPDS cyberinfrastructure or other information systems. We use the CrossReferences property set for other references concerning the resource entity on other websites (that are not necessarily part of the NPDS cyberinfrastructure), such as with a DOI for published literature articles or links to publishers' web pages for journal articles. Those metadata properties required in the form of URIs and URLs can be quality checked by parsing them through a regex filter in order to ensure they comply with the requirements for such URIs and URLs.

PORTAL metadata properties that allow free-form text, such as the EntityName, EntityNature, SupportingTags, and OtherTexts may have more or less strict requirements to support flexibility as desired by administrators of domain-specific problem-oriented repositories. Taken as a whole, these free-form property sets can be evaluated for coherence where all sets match concordantly in terms of topic described, calculated using semantic distance between sets [22]. EntityName, EntityNature, and SupportingTags can be quality checked by ensuring that they comply with the concept validity requirements of key terms as required by the particular diristry, where denoting an item as concept valid implies that the particular item has passed the quality check for that set. For the domain-specific field of biomedical artificial intelligence, as an example with the Beacon diristry, terms in the MeSH thesaurus [23] are given which relate to that domain such as "biologic" and "machine intelligence". These property sets are then checked for whether they contain terms described in the thesaurus. The OtherTexts metadata property remains available for flexible use by the repository administrator and/or users for a variety of diverse purposes from self-description of NPDS servers to storage of bibliographic citation metadata in the form of BibTeX/XML [14]. The administrator of the particular repository may impose further requirements in terms of required, permitted, and optional XML tags where the item is parsed to ensure that it has valid XML and that the tag is syntactically checked with lexical parsing. An example of this approach uses an enum for which the required item is the minimal bibliographic citation, while the permitted item is the maximal bibliographic citation with all metadata including a copy of the reference's abstract.

DOORS Metadata Quality Checks

As defined by the original separation of concerns in PDP [5], while the PORTAL metadata property sets remain lexical, the DOORS metadata property sets must comply with semantic standards. All DOORS metadata property sets can thus go through syntax format parsing in order to validate correct RDF, OWL, XML, JSON, HTML, or XHTML. The desired syntax format can be in the form of a dropdownlist menu from which the user selects their preferred format with values given as an enum declared by the administrator of the NDPS service. In this way, we can ensure that DOORS property sets remain compatible with the semantic web. An example of a record with all DOORS property sets containing content is illustrated in Figure 6, where some PORTAL property sets remain empty but the metadata record qualifies nevertheless as valid on the DOORS Infoset.

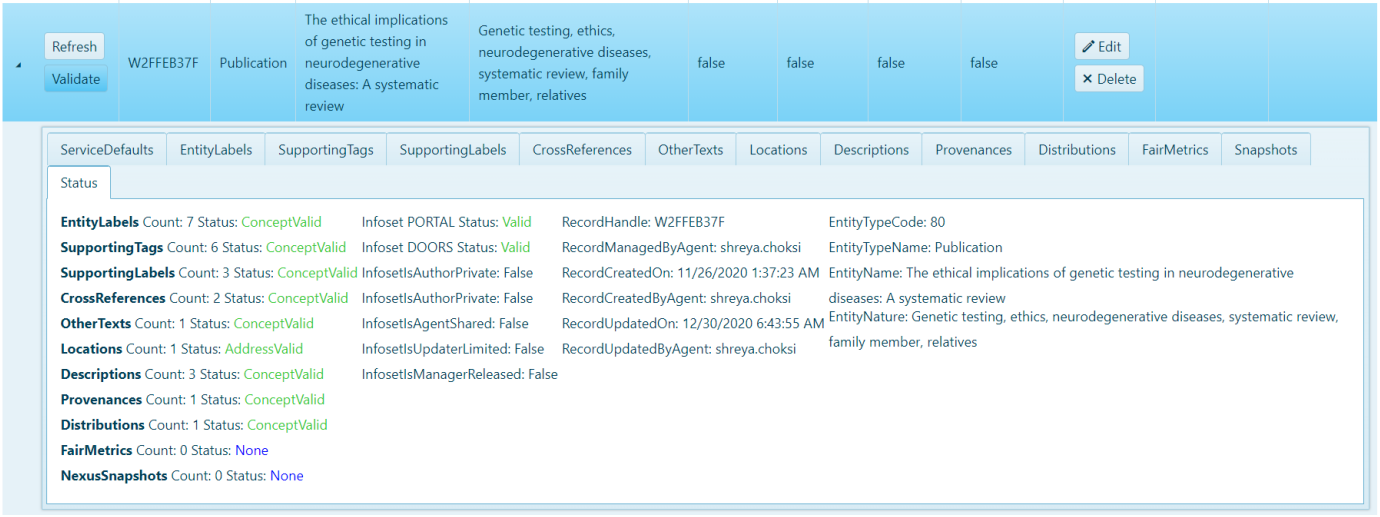

Figure 3: Quality Record from the HELPME Diristry.

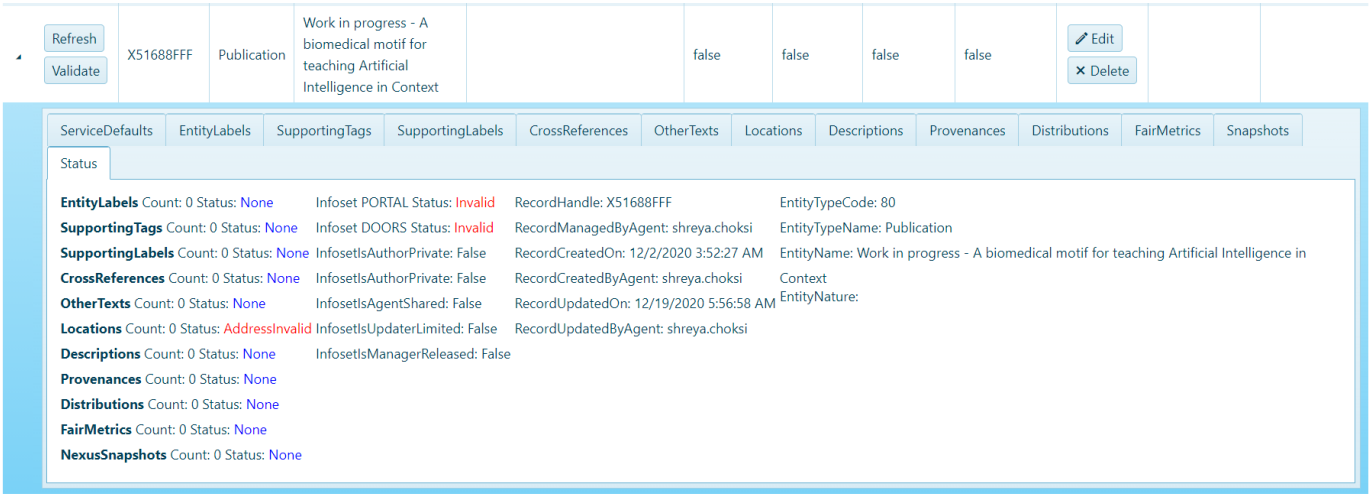

Figure 4: Invalid Record from the HELPME Diristry.

For the DOORS Locations property set, validation checks can assess different kinds of locations including both physical and 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. Absence of response without any returned ping and media type implies that the URL constitutes a dead or broken link and/or some change resulting in failed DNS resolution of the URL. Locations that are postal service mail addresses and geophysical addresses can be validated by a variety of geolocation services (BingMaps is one example) for verification of the normalized mail delivery address as well as the latitude and longitude coordinates. Checking for specified locations of a given resource serves as an indication of the availability and accessibility of that resource [24].

For the DOORS Descriptions property set, 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. This reasoning analysis extends beyond the simpler validation checks for compliance with a syntax standard. Semantic metadata in RDF form may have serialization format specified, where serialization refers to the "representation of RDF data in some RDF serialization format such as Turtle, RDF/XML, JSON-LD, N3, N triples, N quads, or Trig" [25]. The reasoning agents applied to this metadata may extract two sets of numbers with regard to numbers of triples and numbers of inferences from those claims, where these numbers can then be displayed as quantitative descriptors for quality of the content claims. Administrators of the diristry may impose requirements on minimum numbers of claims and inferences where certain values may be below the threshold and marked as bad quality while others may be above the threshold and marked as good quality. These descriptors will serve as a measure of the complexity or 'richness' of the semantic metadata associated with the resource, where a greater number of logical inferences from RDF triples implies 'richer' metadata. In addition to validation of format, correct usage of vocabularies can also be taken into account [26]. Quality of attached RDF metadata can be further evaluated using quantitative metrics. Behkamal *et al.* [27] describe several, which include evaluating syntactic accuracy of documents, uniqueness of the dataset, and consistency and completeness. As an example, syntactic accuracy may be described using the ratio of syntactically incorrect triples, while other ratios may be evaluated as the number of triples with improper properties over

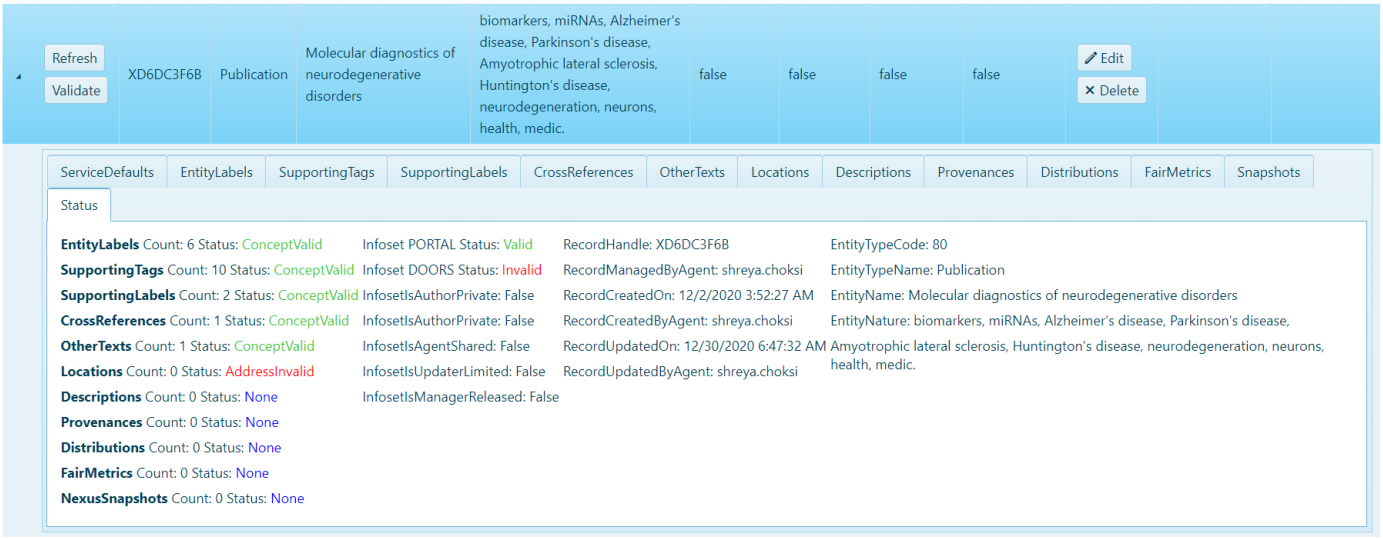

Figure 5: PORTAL Infoset Valid Record from the HELPME Diristry.

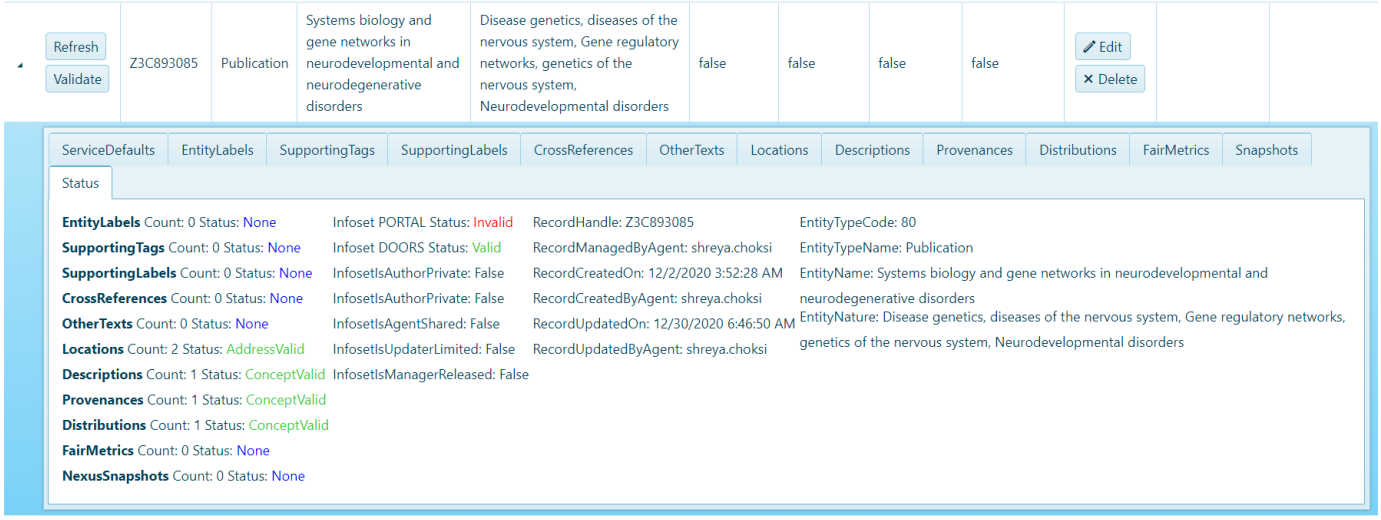

Figure 6: DOORS Infoset Valid Record from the HELPME Diristry.

the total number of triples [27]. Additional related descriptors can also be used for describing incomplete, duplicate, and ambiguous instances along with inaccurate labeling and classification of objects in semantic metadata [28].

The DOORS Provenances property set may document cited references for the resource entity and/or its sources, origins or tracking through versions and/or owners. Cited references can be in the form of all references for a given published article in the scientific literature. Origins may refer to chain of ownership or custody for the given resource. In the world of art, such provenance could be in the form of the sequence of owners for a painting (in which the painting is the resource), where the owners beginning with the artist who created the work, then the subsequent owners who bought and sold the work of art. In criminology, chain of custody may describe a record of how material evidence changed hands during a criminal investigation and trial [29], and be documented in the form of a chronological 'paper trail' which refers to documentation for the sequential order in which individuals handle or collect a certain piece of material evidence [30]. As such, there are a variety of different approaches by which quality checks can be implemented for DOORS Provenances. In the case of bibliographies with cited references for those entities that represent publications in the literature, quality checks may involve verifying those cited sources. Citation elements such as reference type, author, volume, issue, *etc.* that are tagged in XML can be used to construct search queries to external databases like PubMed and CrossRef, where the citation is validated if a match can be made. For a chain of custody or ownership, items in the Provenances set can be checked for whether the XML wrapper tags define persons or organizations to whom the resource belongs. Other characteristics regarding the quality of the resource entity and its metadata can also be derived from the Provenance information. Source reputation "relates the quality of data with the perceived trustworthiness of its source" where a preference or trust order can be established using numerical values for each source [31]. Freshness can relate the "quality of data with its recency" depending on how old the source is [31]. For entities representing datasets, creation and version history can also be tracked, with descriptors of timestamps for creation and updates, as well as version information [32].

The DOORS Distributions property set remains an analogy to the DNS system for which distributions refer to where the mobile metadata of the DNS records are both distributed and redistributed [6], [33]. In addition, the Distributions set may also provide details on both licensing, permissions, and the extent of distribution and/or any restrictions on the distribution. Possibilities include specifications on who can use the particular resource and metadata about the resource, as well as the type of license. Licenses can be both machine-readable and human-readable [34], [35]. In terms of distributions, items may specify where the resource may be distributed and redistributed and what are the constraints and limitations on both distribution and redistribution of the resource. Quality checks can involve parsing for wrapper tags that contain the term "license" and have attributes or properties that refer to permissions, conditions, and limitations of the license. Distribution can be specified with wrapper tags that contain the term distribution or specify expiration, and content can then be checked for whether it refers to a range of locations or numerical value for duration of availability.

Tables 1 and 2 summarize a variety of these quality checks comprised of a combined collection of indicators, quantitative descriptors, and categorical descriptors. General quality checks applying to all metadata property sets consist of I_P as an indicator describing presence of absence of an element in a metadata property set as well as Q_E denoting integer counts of these elements. Quality checks pertaining to specific metadata property sets include examples such as Q_T for integer counts of the numbers of RDF triples and inferences, and C_S for categorical values for syntax and serialization formats. As these examples demonstrate, descriptors and measures of the data unique to each of the named metadata property sets 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 metadata property sets.

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" [2]–[4]. It thus remains important to promote and maintain quality in information systems that deal with the exchange of data and metadata. NPDSLINKS is our approach from the PORTAL-DOORS Project (PDP) which combines our LINKS algorithms and applications with the NPDS cyberinfrastructure through its registries, directories, diristries, and registrars, aimed towards the storage, retrieval, and transfer of data and metadata via a distributed system of network repositories. As a learning intelligence and knowledge system, NPDSLINKS remains equipped to deal with problems of information sharing and will be used for brain health and other biomedical applications. In this report, we outlined some of our plans with 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.* [13] with automated meta-analyses of the clinical trial literature. To follow our ongoing progress with PDP on the NPDS cyberinfrastructure and LINKS applications, visit our new web sites at www.NPDSLINKS.net and www.NPDSLINKS.org.

Citation

Shreya Choksi and Carl Taswell, "The Nexus-PORTAL-DOORS-Scribe (NPDS) Learning Intelligence aNd Knowledge System (LINKS)"; Brainiacs Journal 2020, Volume 1, Issue 1, Edoc B61CA3D89, Pages 1–9; received 2020-Dec-07, published 2020-Dec-30.
Correspondence: CTaswell at Brain Health Alliance
URL: www.BrainiacsJournal.org/arc/pub/Choksi2020LINKS
DOI: 10.48085/B61CA3D89

References

[1] S. Choksi, P. Hong, S. Mashkoor, *et al.*, "NPDSLINKS: Nexus-PORTAL-DOORS-Scribe Learning Intelligence aNd Knowledge System," in *2020 Second International Conference on Transdisciplinary AI (TransAI)*, IEEE, Sep. 2020. DOI: 10.1109/transai49837.2020.00027.

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

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

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

[5] 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.

[6] ——, "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. URL: www.mdpi.com/1999-5903/2/2/156/.

[7] ——, "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.

[8] 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. URL: www.iiisci.org/journal/sci/FullText.asp?id=BA947YN17.

[9] 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.

[10] 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. URL: www.portaldoors.org/pub/docs/ECAI2019DREAMFAIR0612.pdf.

[11] A. Athreya, S. K. Taswell, S. Mashkoor, *et al.*, "The essential enquiry 'equal or equivalent entities?' about two things as same, similar, related, or different," *Brainiacs Journal of Brain Imaging And Computing Sciences*, vol. 1, PEDADC885, pp. 1–7, 1 Dec. 30, 2020.

[12] 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. URL: www.portaldoors.org/pub/docs/ECAI2019DREAMFAIR0612.pdf.

[13] 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. URL: www.portaldoors.org/pub/docs/WRSNM2015TnT1p1020.pdf.

[14] S. K. Taswell, K. Uhegbu, S. Mashkoor, *et al.*, "Storing bibliographic data in multiple formats with the NPDS cyberinfrastructure," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, Oct. 2020. DOI: 10.1002/pra2.428.

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

[16] L. Razmerita, A. Angehrn, and A. Maedche, "Ontology-based user modeling for knowledge management systems," pp. 213–217, 2003. DOI: 10.1007/3-540-44963-9_29.

[17] M.-A. T. Vu, T. Adalı, 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.

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

[19] 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.

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

[21] 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. URL: www.portaldoors.org/pub/docs/ASIST2019FairMetrics0611.pdf.

[22] X. Ochoa and E. Duval, "Quality metrics for learning object metadata," in *EdMedia+ Innovate Learning*, Association for the Advancement of Computing in Education (AACE), 2006, pp. 1004–1011.

[23] C. Taswell, G. TeleGenetics, and C. Ladera Ranch, "Use of the mesh thesaurus in the portal-doors system," in *Proceedings AMIA 2010 Symposium on Clinical Research Informatics, San Francisco CA*, 2010, AMIA–033.

[24] A. Zaveri, A. Rula, A. Maurino, *et al.*, "Quality assessment for linked data: a survey," *Semantic Web*, vol. 7, no. 1, P. Hitzler, Ed., pp. 63–93, Mar. 2015, ISSN: 22104968, 15700844. DOI: 10.3233/SW-150175.

[25] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, *et al.*, "A comprehensive quality model for linked data," *Semantic Web*, vol. 9, no. 1, A. Zaveri, D. Kontokostas, S. Hellmann, *et al.*, Eds., pp. 3–24, Nov. 2017, ISSN: 22104968, 15700844. DOI: 10.3233/SW-170267.

[26] M. Knuth, D. Kontokostas, and H. Sack, "Linked data quality: Identifying and tackling the key challenges," vol. 1215, Sep. 2014.

[27] B. Behkamal, M. Kahani, E. Bagheri, *et al.*, "A metrics-driven approach for quality assessment of linked open data," *Journal of theoretical and applied electronic commerce research*, vol. 9, no. 2, pp. 11–12, Aug. 2014. DOI: http://dx.doi.org/10.4067/S0718-18762014000200006.

[28] Y. Lei, V. Uren, and E. Motta, "A framework for evaluating semantic metadata," 2007. DOI: https://doi.org/10.1145/1298406.1298431.

[29] A. Badiye, *Chain of custody*, Sep. 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK551677/.

[30] R. Longley. (Nov. 3, 2020). "What is chain of custody? definition and examples," URL: https://www.thoughtco.com/chain-of-custody-4589132.

[31] G. Flouris, Y. Roussakis, M. Poveda-Villalón, *et al.*, "Using provenance for quality assessment and repair in linked open data.," in *EvoDyn@ISWC*, T. Groza, D. Plexousakis, and V. Novácek, Eds., ser. CEUR Workshop Proceedings, vol. 890, CEUR-WS.org, 2012. URL: http://dblp.uni-trier.de/db/conf/semweb/evodyn2012.html#FlourisRPMF12.

[32] A. Assaf, R. Troncy, and A. Senart, "Roomba: An extensible framework to validate and build dataset profiles," *CEUR Workshop Proceedings*, vol. 1362, pp. 32–46, Jan. 2015.

[33] 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.

[34] J. Debattista, C. Lange, S. Auer, *et al.*, "Evaluating the quality of the lod cloud: An empirical investigation," *Semantic Web*, vol. 9, no. 6, R. Verborgh, Ed., pp. 859–901, Sep. 2018, ISSN: 22104968, 15700844. DOI: 10.3233/SW-180306.

[35] M. Ben Ellefi, Z. Bellahsene, J. Breslin, *et al.*, "Rdf dataset profiling - a survey of features, methods, vocabularies and applications," *Semantic Web*, vol. 9, Aug. 2017. DOI: 10.3233/SW-180294.