Building a Web of Linked Data Resources to Advance Neuroscience Research

Authors
B Nolan Nichols[1,2]*
Satrajit S. Ghosh[3,4]*
Tibor Auer[5]
Thomas Grabowski[6]
Camille Maumet[7]
David Keator[8]
Kilian M. Pohl[1,2]
Jean-Baptiste Poline[9]

*co-primary authorship

Affiliations
[1]Center for Health Sciences, SRI International, Menlo Park, CA
[2]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA
[3]McGovern Institute For Brain Research, Massachusetts Institute of Technology, Cambridge, MA
[4]Department of Otolaryngology, Harvard Medical School, Boston, MA
[5]MRC Cognition and Brain Sciences Unit, Cambridge, UK
[6]Department of Radiology, University of Washington, Seattle, WA
[7]Warwick Manufacturing Group, University of Warwick, Coventry, UK.
[8]Department of Psychiatry and Human Behavior, University of California, Irvine.
[9]Helen Wills Neuroscience Institute, H. Wheeler Brain Imaging Center, University of California Berkeley.

Abstract
The fundamental goal of neuroscience is to understand the nervous system at all levels of description, from molecular components to behavior. An emerging set of Web technologies, known as Linked Data, is changing how data and knowledge about the nervous system can be accessed and used. This paper provides an introduction to these technologies and reviews their application and influence on neuroscientific research.

Introduction

Scientific progress depends on the ability of researchers to generate, support, or refute hypotheses about theories based on systematic observations and rigorously-acquired data. The complex interaction between new data-driven disciplines, efforts to reproduce results, and economic pressures is creating a paradigm shift in this scientific process. First, the discipline of "data science" has advanced the scientific community's fundamental understanding of the potential for computational methods and cyberinfrastructure to accelerate biomedical knowledge discovery[1]. Second, in

recognizing that certain results are not as reproducible as previously thought[2–6], the "reproducibility crisis" has motivated new efforts to remedy frail discoveries, such as those caused by inadequate data documentation[7,8]. Finally, economic pressures are driving society and funding agencies to demand more cost effective and quicker translation from basic science into clinical practice by broadening data sharing and reuse efforts.

In response to these demands, large-scale research initiatives are being deployed worldwide[9,10]. For example, the Office of Data Science (ODS) at the National Institutes of Health (NIH) implemented the Big Data to Knowledge (BD2K) initiative[1] and the European Life-science Infrastructure for Biological Information (ELIXIR)[11] has been deployed to integrate life science information. In neuroscience, the NIH Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative[12] and European Union Human Brain Project[13] are developing data science infrastructure in partnership with community efforts, such as Research Data Sharing Without Borders[14]. By developing data science communities, these and similar efforts[15,16] aim to elicit a paradigm shift from individual laboratory-based work, towards a more open and collaborative scientific community that produces Findable, Accessible, Interoperable, and Reusable (FAIR)[17] research deliverables. The scope of this vision spans analytical tools and computational infrastructures, including efforts for better data management and integration.

Current challenges to effective data management and integration include the prodigious rate by which digital neuroscientific data are being generated, the diversity of the data (from molecules to systems across species), and the multitude of analyses software that process the data. For example, the BigBrain effort constructed a histological human brain atlas containing 1 terabyte of multi-modal data[18] and a 0.3-cubic-centimeter of a single mouse brain processed using CLARITY results in 4.8 terabytes of raw data[19]. However, much of the data produced worldwide remains inaccessible. A general inaccessibility of data[20],[20]due to a lack of integration within and across laboratories, has undermined the neuroscience community's ability to optimally translate data into knowledge. One possible solution to this issue developed by data science[21] is to include precise metadata (i.e., descriptions of data in a computer accessible manner). An example of metadata is "provenance" (i.e., information that chronicles how data were generated and processed) represented through Linked Data, a set of principles that use Web technologies to associate data within and across communities. The remainder of this article will review the concepts and technologies behind this type of metadata focusing on their impact on effective data reuse and scalability of research in neuroscience.

Goals, governing principles, and building blocks for a Web of Linked Data

Research utilizes data and known facts and produces new information, often through an empirical, Popperian approach[22]. The ability to effectively discover and integrate

existing data and facts is therefore crucial to any research finding. As the global scientific output is estimated to double every nine years[23], the risk of missing relevant information is increasing. To harness this deluge of information, scientific resources (e.g., data, analysis tools, and publications) are in the process of being indexed with links to related information (e.g., provenance and source code) that are easily searchable. The Web offers a scalable and decentralized architecture to support the management and retrieval of information. Governing principles have been proposed that clarify how to link data on the Web by leveraging computer-processable metadata, vocabularies that formalize the meaning of metadata terms, and query protocols for retrieving information from distributed sources.

Governing principles for a web of data

Historically, libraries and library science[24] governed access to the organization of information and curated knowledge. Referred to by Ranganathan as "The Five Laws of Library Science," these design principles helped build a library framework that allowed readers to readily find relevant books while recognising that libraries change over time. This accessibility was facilitated by a taxonomical index, but did not link across books or their content. Similar principles are applied to MESH descriptors used by Pubmed[25],[25]may tag a manuscript as a '*primary research article*' using '*functional MRI*' to study '*autism*'. While providing simple descriptions of publications, such metadata lack a rich description of their content (e.g. the facts contained in these publications, or the data used in the study). A richer description and linking of the content would enable machine responses to questions such as, "What manuscripts have data that supports opioid excess theories of autism?" or "What autism studies show activation in limbic structures using task-based fMRI?" Furthermore, study replication and testing of alternate hypothesis could benefit from queries linking back to fine grained experimental conditions (e.g., acquisition parameters), statistical methods (e.g., thresholds or p-values), or to the dataset used for analysis. To enable such queries, metadata that link to each other are necessary. In light of this need for metadata, a growing number of publishers have adopted the use of minimal information sets[26]. For example, the ISA-Tab[27] standard is used by Nature Scientific Data to tag publications with study description that can be represented as Linked Data and queried. To standardize such information, a community organization comprised of academics, non-profits, and industry partners called the Future of Research Communications and e-Scholarship (FORCE11) has proposed the FAIR data principles (see Box 1). FAIR data principles extend the rules behind library curation to the Linked Data space improving governance of data and metadata, such as access to HIPAA regulated patient information. The Web is an ideal platform for implementing these principles, which provide specificity to the more general Linked Data Principles[28] (discussed later) outlined by the World Wide Web Consortium (W3C)[29].

Box 1 FAIR Data Principles

The FAIR data principles, developed by FORCE11, proposes the use of Web standards for organizing metadata to transform how data are (re)used[30].

FORCE11 defined a set of principles that are designed to make data FAIR - Findable, Accessible, Interoperable, and Reusable[17]. Specifically, the principles are below.

To be Findable:
F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be Accessible:
A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:
R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

How did the Web become a scalable infrastructure for organizing data?
As summarized in Box 2, the Web began as a *web of documents, i.e.,* an infrastructure for serving and linking within and across HTML documents that could be accessed and read by people using Web browsers. It then evolved into an open forum where the general public published content. It is now being restructured into a *web of data* that can, programmatically, be reused and repurposed to create more efficient and adaptive applications. In this section we review the key W3C standards and other technologies that allow generation, query, and linking between computer processable information descriptions. We illustrate this with an example of how such standards can represent and enrich information stored in spreadsheets (e.g., Excel), still a very common format for storing and sharing data in neuroscience.

Box 2 Evolution of the Web

Initially, the **Web (1.0)** provided a publishing platform for static documents akin to a digital version of books or newspapers for people to read and discuss. The document creators provided clickable links to other documents, much in the same way authors reference other manuscripts, but with an interactive component that enabled Web "surfing" of what are typically read-only documents.

As the **Web (2.0)** became 'read-write,' even novices and non-computer savvy users were able to contribute content using Web applications that make it simple to blog or comment on the work of others. In addition, pages became dynamic and started to update content without interaction from the reader.

The **Web (3.0)** is becoming a web of data that can be used by people and computers to access and gain knowledge from data and documents. By embedding semantic information about webpage content that computers can process, algorithms are able to automate many tasks, such as recommending content tailored to individual interests and linking related services (e.g., calendar and email). For example, a computer can follow links between Web pages, read information on the page, and display it to a user. Links within and between Web pages can contain hidden tags or labels that provide meaning to specific content (e.g., page title, article author, or related content) that computers can use in predefined ways. In this way, content can be woven and linked to form a web that machines can manipulate in meaningfully.

Spreadsheets, such as in Figure 1A, reduce the complexity of scientific data to a flat (2D) data structure that generally belies the complexity of original datasets (e.g., by providing summary measurements of 3D or 4D observations). They do so by placing significant weight on implicit knowledge or by defining an external document that decodes their columns (e.g., a data dictionary). To automatically process these details, structure is needed captures these additional metadata. We now review specific Web standards for describing data such as, for example, the ones found in spreadsheets.

A Common Data Model: The Resource Description Framework (RDF) is a W3C recommendation for representing information in the Web[31] as a network of resources. At its core, this data model enables a scalable, Web-based graph database. When connected it forms a directed, labeled graph, where each edge represents a specific type of relation or link between two resources (i.e., Web addresses) or between a resource and a value (e.g., an integer, string, etc.). This "node - edge - node" is referred to as a "triple" and a dataset consists of a set of triples. Graph representations offer an intuitive approach for conceptualizing data that can be represented by visualising both the data and their relationships (e.g., as a concept map). This data model can then be used to represent, not only all the information that may be contained in a spreadsheet, but augment it with additional metadata. Furthermore, by using URIs (Uniform Resource Identifiers) as the names for nodes and edges, RDF datasets can link to resources on

the Web. Figure 1B provides an example of how a spreadsheet can be augmented with RDF by attaching additional information that links to content on the Web.

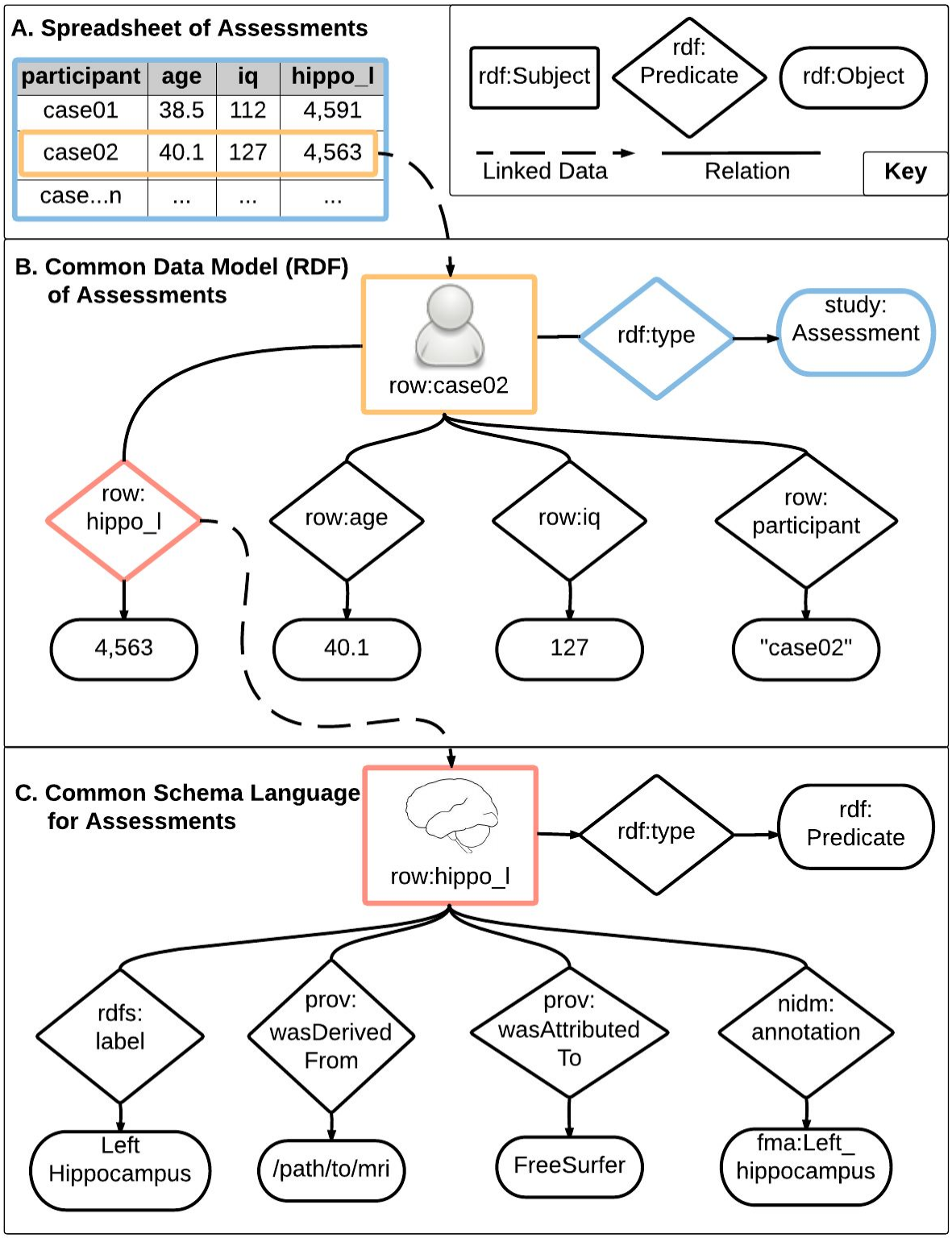

Figure 1. Spreadsheets on the Web. Spreadsheets are a common format for exchanging and processing tabular data in analysis packages. **A)** This spreadsheet of

assessments data contains records (i.e., rows) with columns denoting a participant identifier (participant), demographic information (age), a neuropsychological test result (iq), and an anatomical measurement (hippo_l). **B)** Each cell in this tabular spreadsheet can be represented as an RDF statement (i.e., triple) with a Subject, Predicate, and Object or Value. Each row is assigned a Web address to represent the 'Subject' of a statement. The 'Subject' is denoted here using 'row:case02' where the 'row:' prefix (i.e., namespace) is shorthand for the base Web address (e.g., http://my.study.com/row#) that is expanded from 'row:case02' to 'http://my.study.com/row#case02'. The columns of the spreadsheet are modeled as a 'Predicate' that connects the 'Subject' to a specific 'Object' or 'Value' to form a statement. The 'rdf:type' is used to declare that this row of the spreadsheet is about a 'study:Assessment', while the remaining 'Predicates' connect to specific data values for each column (i.e., participant, age, iq, and hippo_l). **C)** Each 'Predicate' is also denoted using a Web address, thus allowing it to assume the role of a 'Subject' in another set of statements that provide contextual metadata (i.e., a schema). These metadata can be used, for example, to decipher the term 'hippo_l' as "Left Hippocampus", that the measure was derived from an MRI scan, that the analysis was attributed to the FreeSurfer software, and that it is annotated with the URI 'fma:Left_Hippocampus' from the Foundational Model of Anatomy (FMA) ontology[32,33]. Using the RDF representation, both data and metadata can be stored in the same document or distributed across the Web. Represented using the common data model, this data contains links that facilitates information retrieval and integration tasks through additional documentation.

A Common Schema Language: A schema language provides constructs to describe the organization of data. Schema languages define the expected nodes and edges in a given document that is expressed using a data model (e.g., RDF). Two popular schema languages are RDF Schema (RDFS)[34] and the Web Ontology Language (OWL)[34,35]. Analogous to the use of symbols (e.g., written, verbal, and nonverbal) to communicate concepts in natural human language, these languages use Web "resources" (i.e., addresses on the Internet) as symbols to convey meaning. For example, Neurolex[36], a Linked Data application providing community curated neuroscience knowledge, uses the URI: http://uri.neuinfo.org/nif/nifstd/birnlex_2092 for defining "Alzheimer's disease". These symbols can be organized into an ontology, a formal representation of a set of concepts within a domain and the relationships between those concepts. By organizing concepts by category (e.g., "Neurodegenerative disease"), a vocabulary can be defined and related symbols be discovered. "Cerebral Palsy" and "Parkinson disease" are examples of related symbols that may contain structured links to specific content including definitions, synonyms (e.g., Alzheimer's dementia), or abbreviations (e.g., AD), and included as part of a disease ontology (e.g., the Disease Ontology[37] or Autism Ontology[38])

A Common Query Language: Current data models and schemas aim to improve the efficiency of information retrieval on the Web. While links and shared words across

websites can be centrally indexed (e.g., Google), scientific data will continue to be decentralized and distributed across individual labs. Therefore it becomes necessary to query at the scale of the Web using “federated” approaches that access data sources distributed across many locations. The, recursively named, “SPARQL Protocol And RDF Query Language” was designed to fulfill these requirements and, unlike traditional databases for tabular data, offers Web-accessible endpoints to issue queries against. Similar to SQL, the query language for traditional relational databases, SPARQL uses selection and projection criteria to retrieve a given view of a database. The key difference is that SQL is designed for accessing tabular data while SPARQL uses graph-based pattern matching that enables more flexible queries Like traditional databases, datasets can be updated with new facts, thereby enabling the development of a web of Linked Data.

Common Security Protocols and Provenance: In the context of mental health and genetic information, standards for security and provenance are essential to create a balance between available data and privacy. To attain this balance, security protocols such as WebID, together with CORS (Cross-Origin Resource Sharing) and provenance tracking, can be used to augment the Web with policies for resource sharing[39], including authorization for specific use. PROV is a schema recommended by the W3C for documenting provenance[40]. It defines provenance as “...information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.”[41]. Provenance can be encoded with PROV using the common data model, annotated with existing vocabularies, queried, and connected as Linked Data. The PROV schema has also been extended to accommodate use cases in neuroscience, as exemplified by the Neuroimaging Data Model (NIDM) standard for capturing provenance in brain imaging research[42]. Together with security provisions for authorized access, any unauthorized use can now be tracked. For example, if researcher "A" is allowed to have data but not researcher "B", a secure provenance network can track if "B" used data from "A" and potentially automate resource authorization or denial.

Linked Data: Linked Data comprises three principles (Figure 2). First, Linked Data uses HTTP URIs to name both the nodes and the relationship between nodes (i.e., predicates). The second principle recommends that when an URI is processed (i.e., dereferenced) by a Web browser or other HTTP compatible application, that useful, computer processable, data is provided using Web standards. The final principle states that the data provided contain links to other IRI nodes that, in turn, should contain additional IRIs, thus ultimately creating a *web of data* that allows for the discovery of more resources, for example, with user-friendly Web applications such as LodLive Linked Data Browser[43]. These principles can be realized using the RDF, RDFS/OWL, and SPARQL standards described above.

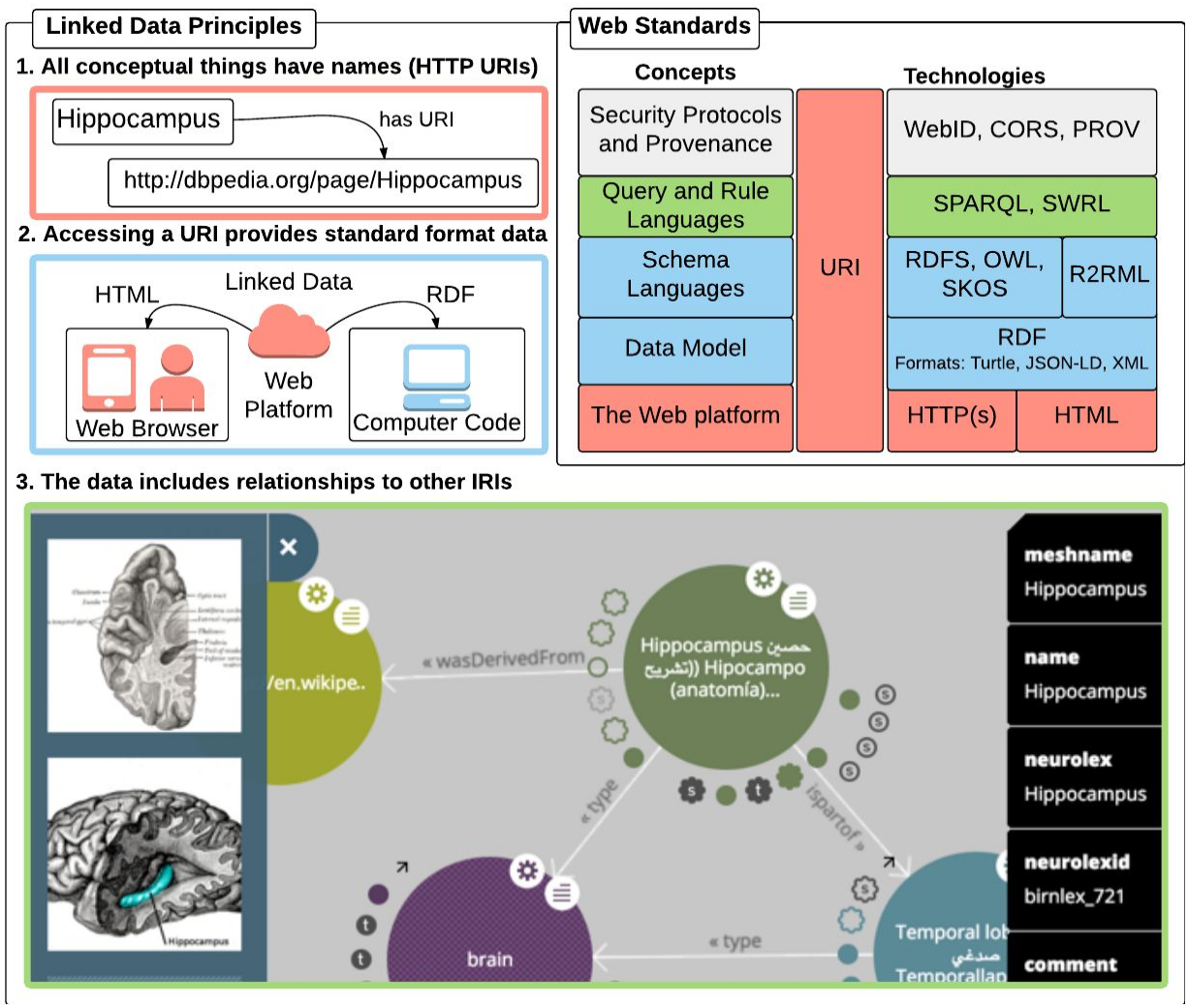

Figure 2. Linked Data Abstraction. The Linked Data Principles [44] comprise a set of three ideas that allow data on the Web to be represented, found, and related. 1) Concepts should be given names using HTTP URIs that 2) can be accessed using a Web browser or programming language that returns data in a standard format, such as HTML or RDF, with 3) related information encoded via links to Web accessible data. These principles can be implemented using a subset of Web standards (Web platform, Data Model, WebID, and CORS) to interconnect data. The graphical representation of the links shown in 3) was generated by the LodLive Linked Data Browser[43].

The application of Linked Data is gaining prominence to organize knowledge. For example, the Google Knowledge Graph[45] uses these technologies to optimize search performance and to enhance display results by encouraging the use of structured data markup (i.e., RDFa) in web pages[46]. DBPedia extracts general information from Wikipedia on people, places, and things and then exposes it as an RDF knowledge

base, a view of which is displayed on Wikipedia[47]. At the British Broadcasting Channel (BBC), New York Times, and Nature Publishing Group (NPG), all article metadata (e.g., titles, authors, and abstracts) are made available as RDF, thus enabling computer processing of this information, and making it a part of the Linked Open Data Web.

In summary, the result of these existing technologies is a Web-based and Web-scalable informatics architecture that facilitates support of the FAIR principles (i.e., to allow data to be findable, accessible, interoperable and reusable through common standards). Using the Linked Data Principles and Web Standards can enable researchers to access and reuse data both internally, across collaborators, and on the Web.

Applications of Linked Data in neuroscience

The technologies and principles described above can, and already are, impacting the field of Neuroscience by providing rich data descriptions and powerful search applications as exemplified in the following.

An emerging ecosystem of digital research objects

Defining metadata relevant for a given domain allows the integration of research data by linking them as digital research objects. The Minimal Information for Biological and Biomedical Investigations (MIBBI) project is an umbrella organization for defining such minimal information standards[26]. The Biosharing effort grew out of MIBBI to offer an enhanced repository of digital resources that includes data standards, databases, and policies across biomedical research[48]. By capturing information about biomedical resources using structured data, Biosharing metadata can be used in search applications and for precise data description. For example, the Center for Enhanced Data Annotation and Retrieval (CEDAR)[49] is developing user-friendly interfaces using templates that assist in complying with minimal information reporting standards. Furthermore, Linked Data vocabularies or ontologies hosted by resources such as BioPortal[50] and Ontobee[51] supply Web services that can support automation of data entry, thus reducing manual effort.

Leveraging bioinformatics resources

There are a growing number of Linked Data resources in the field of bioinformatics that are being effectively leveraged for neuroscience research. Resources that are essential to genomics, such as the Gene Ontology[52] and Entrez Gene[53] [53]a[53]a[53]a[53]e[53]are being represented using Linked Data. Similarly, the Human Phenotype Ontology[54] and Mouse Genome Database[55] contain useful information for comparing across species, and have only recently been integrated using Linked Data approaches. By linking these resources through Linked Data technologies, these disparate bioinformatics resources are becoming a web of data that can be queried as a single, integrated database and provide answers to detailed neuroscience questions.

The Bio2RDF project is an early example of integrating bioinformatics resources. Bio2RDF applies Linked Data technologies to enable queries that were not previously

possible programmatically (i.e., using a query language). Using the Bio2RDF framework, Belleau et al.[56] linked four genes to Parkinson's disease to answer two questions in Parkinson's disease research related to:

1. Which Gene Ontology (GO) terms describe four genes of interest (Rxr, Nurr1, Nur77, and Nor-1)?
2. Which articles mentioning these four genes of interest are related to apoptosis AND cytoplasm?

The Gene Ontology[52] was queried for identifiers of these genes. In a second query, data collected from Entrez Gene[53] and Pubmed[25] sources into Bio2RDF were accessed to identify all articles with keywords and Gene Ontology annotations for "apoptosis and cytoplasm" for the genes of interest. This use case demonstrates how data from heterogeneous sources could interoperate using Linked Data, and how the results of a query could be used to get more precise and complete information on Parkinson's disease from the literature. After identifying candidate genes, a next step may be to search across species for animal models of Parkinson's disease. By using a query language and a specific data schema or ontology, more precise questions can be answered using existing data compared to using conventional search engines that only provide natural language documents.

Another example for Linked Data use in biomedical research is the Monarch Initiative[57]. This computational platform uses Linked Data technologies to compare phenotypes within and across species[58]. The underlying RDF database integrates and aligns cross-species gene, genotype, variant, disease, and phenotype data, such that the platform can identify animal models of human disease through similarity analysis of biochimical models. A user of this system can browse human diseases and access links to related animal models and discover related genes. For example, dystonia (http://monarchinitiative.org/disease/OMIM:612067) has seventeen associated phenotypes, one mouse model, one gene, and one pathway. This provides researcher with a single entry point to a networked of information that previously needed to be generated with significant manual effort.

In summary, the general availability of these bioinformatics resources as Linked Data enables information to be combined more efficiently than manual searches. This is exemplary of the shift towards a digital ecosystem for biomedical data science that can be leveraged in neuroscience.

Linked Data technologies in neuroscience

A growing number of neuroscientific applications employ technologies derived from the principles of Linked Data. An early example is SenseLab, which provides an entire suite of interrelated databases for molecular-level neuroscience[59]. Samwald et al. migrated these molecular data from a traditional relational database to a graph database using a semi-automated approach. This process allowed once-siloed data to integrate with additional resources from the community including the Brain Architecture Management

System[60,61], Gene Ontology[52], Subcellular Anatomy Ontology[62], and UniProt[52,63]. Harmonizing these external knowledge-bases further interconnects and enhances the integration of information that is available for search[64]. Similarly, the NeuroMorpho database applies controlled terminologies for metadata labeling and to enhance the search results of queries for neuronal cell structures[65].

Beyond single databases, aggregations of databases and datasets now use Linked Data to simplify exploration of neuroscientific data. The Neuroscience Information Framework (NIF)[65,66] developed technology to crawl neuroscience databases[67] and create a searchable archive of information obtained from controlled vocabularies and ontologies[65,66,68]. More recently, SciCrunch[69] is generalizing the NIF technologies for the broader biomedical domain[69], illustrating how these informatics solutions can reach beyond their initial domain. The National Database for Autism Research (NDAR)[69,70] employs an ontology describing the concepts studied in autism research[38,69,70]. By incorporating the conceptual framework of an Autism ontology into NDAR, specific cognitive measures (e.g., verbal IQ) and their value ranges can be ascribed to a given concept (e.g., low verbal IQ). When rolled out over the entire resource, this enables user to browse through data conceptually, for example, by selecting all subjects labeled with 'low verbal IQ.'

Another potential of Linked Data is the ability to retrieve information computationally that would otherwise require specific studies or significant manual effort. For example, Poldrack *et al.*[71] used a clustering algorithm to define "topics" in cognitive functions and in disorders using the NeuroSynth database[72], a collection of brain imaging result coordinates and associated terms from publications. The associations between cognitive functions and disorders were determined using an automated similarity analysis of the brain images associated with these topics. Similar inferences can be drawn from the "Linked Neuron Data" (LND) resource[73], which provides a large aggregation of Linked Data that includes facts similar to those used by Poldrack et al. Fourteen of eighteen relations from the top three associations inferred in Poldrack et al., could be recovered using queries on the LND resource, highlighting the use of a Linked Data framework as an alternative for knowledge representation and inference. See Box 3 for more details on the methods.

Box 3: Use of linked data resource example

We examined the "Linked Neuron Data" (LND, http://www.linked-neuron-data.org) resource and asked the question whether results obtained by Poldrack et al 2012 could be found using this linked data technology resource.

Poldrack et al (2012) first used a topic model to extract topics (a set of words occurring often with each others) from the neurosynth database, both for mental concepts and for disorders. For each of these topics, brain maps were constructed by testing at each voxel if more activity was found in papers including these terms than

for other terms, constructing neural activation maps per topics. The images obtained from the mental concepts and disorders were linked using sparse canonical correlation analysis (sCCA), yielding a number of components linking mental concepts and disorder, with weights on the mental concepts and disorder topics. We found that disorders found by sCCA components best describing the relations between mental concepts and disorders could in part be directly obtained using the LBD *"Cognitive Functions" and "Brain diseases"* relations which are constructed with co-occurrence of terms in documents. For the first three components (canonical variates) of the sCCA, disorder terms found associated with mental constructs could be equated -or approximately equated- to LBD brain diseases in about 14 over 18 of the cases. The correspondence between terms used in (Poldrack 2012) and the LBD was in general possible but not always.

Challenges for the application of linked open data in neuroscience

Vocabulary creation and maintenance

A common and precise vocabulary is necessary to assure data are linked. The development of these dictionaries is at the core of the linked data enterprise and vision (see section on "A common schema language"). Two challenges can be distinguished, with regards to how these vocabularies are maintained. First, there are technical challenges in constructing the tools that allow the creation, curation, search and use of these vocabularies. To ensure that a concept or an entity - described with a simple word, eg "hippocampus", or with a set of words (e.g., "medial temporal lobe structure"), has a specific meaning, a unique identifier must be attributed to each concept or entity. Second, there are social or community challenges. As these vocabularies and resources develop, several versions of the same entity or concept can be proposed, and vocabulary sets overlap, possibly with slightly different version of the same concept. The stewardship of these vocabularies needs to be established and agreed upon by research communities, and there is no broadly agreed upon model for how to do this. Funding agencies and institutions have a key role in maintaining access to these resources.

Extending Linked Data principles to the provenance of data processing

Several publications highlight the "reproducibility crisis" in biomedical research[3,4,6] including neuroscience[2,5]. From psychology[2] to pre-clinical studies[74], it is clear that the neuroscientific community must focus efforts to produce robust and reproducible results. While reproducibility can be defined along a spectrum[75], it generally requires that results obtained from one laboratory can be reproduced in another. This, in turn, requires linking information across data acquisition, processing, and results. Using provenance models, such as PROV, meets the challenge of capturing information associated with acquisition and analysis when with the Linked Data principles. For instance, when a specific version of a neuroimaging tool (e.g., SPM[76], FSL[77], or FreeSurfer[78,79] is used to produce derived data, information about the software, its inputs (e.g., data and parameters) and outputs needs to be recorded for reproducibility. Given the mixture of

automated and manual processing required in many neuroscience experiments, such recording is challenging. Developing or augmenting tools to auto-document such processes will require significant effort. However, the lack of any incentives for publishing reproducible research in many areas of neuroscience makes broad adoption of detailed curation efforts unlikely until the benefits are clearly identified. As an example in brain imaging, the "Committee on Best Practices in Data Analysis and Sharing" of the Organization for the Human Brain Mapping has recommended a minimal set of concepts and information that should be documented[80]. Linked Data technologies can help implement recommendations, but will need to be integrated in research tools and laboratory workflows.

Governing sharing and reuse of Linked Data

There are mature Linked Data software tools for data sharing and knowledge integration. However, in the context of neuroscience and biomedical research on human subjects, it is important to consider and maintain the privacy and confidentiality of study participants. The availability of linked genetic, clinical, behavioral, and imaging data increases the need to use security technologies that address the potential misuse. Current data sharing policies and legal constructs vary in their ability to protect an individual, a family, or a group across countries. With the increased availability of open data, simple acknowledgment of data use terms will be insufficient. To provide trust and allow reproducibility, it will be necessary to maintain a link between published results, including outcomes of clinical trials, and the data that were used. On the other hand, such data, in biomedical research, often contain sensitive information. Thus, governance and legal policies have to evolve together with data sharing technologies. In this context, efforts such as portable legal consent[81] [1] are important and needs to be part of a global conversation on privacy, data use, and reuse.

These challenges need to be perceived as opportunities for collaboration and innovation that need to be addressed by a multidisciplinary neuroscience community.

Conclusion

The rapid proliferation of neuroscientific data and computation leads to a compelling need for standards and platforms to share, access, query, and establish trust in data. These standards and platforms can empower researchers with resources that are more efficient to use, rather than settling for suboptimal cyberinfrastructure to support their research. Web technologies provide the cyberinfrastructure necessary to support the principles of FAIR and Linked Data. These technologies are enabling the scalable tools and services that link across laboratories and research centers. Many such tools are already being used or are emerging in biomedical, and more specifically neuroscience, research. There remain technological and societal challenges that, once overcome, will likely improve scientific productivity and expedite the translation of knowledge from laboratories to therapies and treatments. The web of linked data is likely to be a

[1] see technologies such as HTTPa/i[82,83] and IPFS[84]

disruptive technology for science, and specifically for more rapid and trusted neuroscience discoveries.

AD	Alzheimer's Disease
BD2K	Big Data to Knowledge
BRAIN	Brain Research through Advancing Innovative Neurotechnologies
CEDAR	Center for Enhanced Data Annotation and Retrieval
CLARITY	Clear Lipid-exchanged Acrylamide-hybridized Rigid Imaging-compatible Tissue-hYdrogel
CORS	Cross-Origin Resource Sharing
ELIXIR	European Life-science Infrastructure for Biological Information
FAIR	Findable, Accessible, Interoperable, and Reusable
fMRI	functional Magnetic Resonance Imaging
FORCE11	Future of Research Communications and e-Scholarship
FMA	Foundational Model of Anatomy
FSL	FMRIB Software Library
GO	Gene Ontology
HTML	Hyper-Text Markup Language
HTTP(s)	Hyper-Text Transport Protocol (secure)
IRI	Internationalized Resource Identifiers
JSON-LD	JavaScript Object Notation for Linked Data
LND	Linked Neuron Data
LOD	Linked Open Data
MESH	Medical Subject Headings
MIBBI	Minimal Information for Biological and Biomedical Investigations
NDAR	National Database for Autism Research
NIDM	Neuroimaging Data Model
NIF	Neuroscience Information Framework
NIH	National Institutes of Health
ODS	Office of Data Science
OWL	Web Ontology Language
PROV	W3C Specification for Provenance Information on the Web
SPARQL	SPARQL Protocol And RDF Query Language
SPM	Statistical Parametric Mapping
SQL	Structured Query Language
R2RML	Relational to RDF Markup Language
RDFa	Resource Description Framework in Attributes
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
Turtle	Terse RDF Triple Language
sCCA	sparse Canonical Correlation Analysis
SKOS	Simple Knowledge Organization System
SWRL	Semantic Web Rule Language
URL	Uniform Resource Locator
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WebID	WebID Specification
WWW	World Wide Web
XML	eXtensible Markup Language

Table 1. Acronyms

Acknowledgements

We thank James Brinkley, Rosemary Fama, Maryann Martone, Stephanie Sassoon, Edith Sullivan for their insightful comments on early versions of this manuscript.

SG was partially supported by NIH grants 1R01EB020740-01A1, 3-R01-MH092380-04S2, and 1U01MH108168-01. BNN and KMP were supported in part by NIH grants AA012388, DA041123, HL127661, AA021697, the BD2K supplement AA021697-04S1 01, the Creative and Novel Ideas in HIV Research (CNIHR) Program through a supplement to the University of Alabama at Birmingham Center For AIDS Research funding (NIH P30 AI027767), which is a collaborative effort of the Office of AIDS Research, the National Institute of Allergy and Infectious Diseases, and the International AIDS Society. T.A. was supported by the Medical Research Council (United Kingdom) [MC-A060-53114]. CM has been supported by the Wellcome Trust.

Contributions

BNN, SG, and JBP designed the overall message of the article and together with KMP developed a first draft. CM, TA, DK provided input on the manuscript.

Conflicts

None

References

1. Bourne, P. E. *et al.* The NIH Big Data to Knowledge (BD2K) initiative. *J. Am. Med. Inform. Assoc.* **22,** 1114–1114 (2015).
2. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14,** 365–376 (2013).
3. Ioannidis, J. P. A. How to make more published research true. *PLoS Med.* **11,** e1001747 (2014).
4. Ioannidis, J. P. A. Excess significance bias in the literature on brain volume abnormalities. *Arch. Gen. Psychiatry* **68,** 773–780 (2011).
5. David, S. P. *et al.* Potential reporting bias in fMRI studies of the brain. *PLoS One* **8,** e70104 (2013).

6. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2,** e124 (2005).
7. Carp, J. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* **63,** 289–300 (2012).
8. Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6,** 149 (2012).
9. Pilat, D. & Fukasaku, Y. OECD principles and guidelines for access to research data from public funding. *Data Science Journal* **6,** OD4–OD11 (2007).
10. of Health, U. S. N. I. & Others. NIH Data Sharing Policy and Implementation Guidance. *NIH, Bethesda, MD* (2003).
11. Crosswell, L. C. & Thornton, J. M. ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.* (2012). at <http://www.cell.com/trends/biotechnology/pdf/S0167-7799(12)00017-0.pdf>
12. Jorgenson, L. A. *et al.* The BRAIN Initiative: developing technology to catalyse neuroscience discovery. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20140164–20140164 (2015).
13. Markram, H. The human brain project. *Sci. Am.* **306,** 50–55 (2012).
14. Infrastructure for Understanding the Human Brain. *RDA* (2015). at <https://rd-alliance.org/plenary-meetings/sixth-plenary/programme/e-infrastructures-rda-data-intensive-science/infrastructure>
15. Sherif, T. *et al.* CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front. Neuroinform.* **8,** 54 (2014).
16. Frisoni, G. B. Neugrid: A european infrastructure for scan analysis. *Alzheimers. Dement.* **11,** P121 (2015).

17. The FAIR Data Principles. *FORCE11* (2014). at <https://www.force11.org/group/fairgroup/fairprinciples>
18. Amunts, K. *et al.* BigBrain: an ultrahigh-resolution 3D human brain model. *Science* **340,** 1472–1475 (2013).
19. Tomer, R., Ye, L., Hsueh, B. & Deisseroth, K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. *Nat. Protoc.* **9,** 1682–1697 (2014).
20. Poline, J.-B. *et al.* Data sharing in neuroimaging research. *Front. Neuroinform.* **6,** 9–9 (2012).
21. Margolis, R. *et al.* The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc.* **21,** 957–958 (2014).
22. Popper, K. *Conjectures and refutations: The growth of scientific knowledge*. (routledge, 2014).
23. Bornmann, L. & Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J Assn Inf Sci Tec* **66,** 2215–2222 (2015).
24. Ranganathan, S. R. *The Five Laws of Library Science*. (Ess Ess Publications, 2006).
25. Colaianni, L. A. in *Libraries without Limits: Changing Needs — Changing Roles* 87–92 (1999).
26. Taylor, C., Field, D., Sansone, S. & Aerts, J. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature* (2008). at <http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2008Natur.456..773D&link_type=ABSTRACT>
27. González-Beltrán, A., Maguire, E., Sansone, S.-A. & Rocca-Serra, P. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics* **15 Suppl 14,** S4

(2014).

28. Website. at <http://www.w3.org/DesignIssues/LinkedData.html>
29. Design Issues for the World Wide Web. at <http://www.w3.org/DesignIssues/Overview.html>
30. FORCE11 Manifesto. *FORCE11* (2015). at <https://www.force11.org/about/manifesto>
31. RDF Primer. (2004). at <http://www.w3.org/TR/rdf-primer/>
32. Nichols, B. N. *et al.* Neuroanatomical domain of the foundational model of anatomy ontology. *J. Biomed. Semantics* **5,** 1 (2014).
33. Rosse, C. & Mejino, J. L. V., Jr. The foundational model of anatomy ontology. *Anatomy Ontologies for Bioinformatics* 59–117 (2008).
34. RDF Schema 1.1. at <http://www.w3.org/TR/rdf-schema/>
35. OWL 2 Web Ontology Language Primer (Second Edition). at <http://www.w3.org/TR/owl2-primer/>
36. Larson, S. D. & Martone, M. E. NeuroLex.org: an online framework for neuroscience knowledge. *Front. Neuroinform.* **7,** 18 (2013).
37. Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* **40,** D940–6 (2012).
38. McCray, A. T., Trevvett, P. & Frost, H. R. Modeling the autism spectrum disorder phenotype. *Neuroinformatics* **12,** 291–305 (2014).
39. Seneviratne, O., Kagal, L. & Berners-Lee, T. *Policy-aware content reuse on the web*. (Springer, 2009).
40. Moreau, L., Groth, P., Cheney, J., Lebo, T. & Miles, S. The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web* (2015). at <http://linkinghub.elsevier.com/retrieve/pii/S1570826815000177>

41. PROV Model Primer. at <http://www.w3.org/TR/prov-primer/>

42. Keator, D. B. *et al.* Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* **82,** 647–661 (2013).

43. LodLive - browsing the Web of Data. at <http://en.lodlive.it/>

44. Bizer, C., Heath, T. & Berners-Lee, T. Linked data: Principles and state of the art. in *World Wide Web Conference* 1–40 (2008).

45. Singhal, A. Introducing the knowledge graph: things, not strings. *Official Google Blog, May* (2012).

46. Promote Your Content with Structured Data Markup. *Google Developers* at <https://developers.google.com/structured-data/>

47. Auer, S. *et al.* in *The Semantic Web* 722–735 (Springer Berlin Heidelberg, 2007).

48. Sansone, S.-A. *et al.* Toward interoperable bioscience data. *Nat. Genet.* **44,** 121–126 (2012).

49. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc.* **22,** 1148–1152 (2015).

50. Noy, N. F. *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* **37,** W170–3 (2009).

51. Xiang, Z., Mungall, C., Ruttenberg, A. & He, Y. Ontobee: A Linked Data Server and Browser for Ontology Terms. in *ICBO* (researchgate.net, 2011). at <http://www.researchgate.net/profile/Alan_Ruttenberg/publication/252068564_Ontobee_A_Linked_Data_Server_and_Browser_for_Ontology_Terms/links/0deec5393d9ae86155000000.pdf>

52. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

53. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33,** D54–8 (2005).

54. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83,** 610–615 (2008).

55. Blake, J. A. *et al.* The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39,** D842–8 (2011).

56. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. & Morissette, J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* **41,** (2008).

57. Welcome to Monarch. at <http://monarchinitiative.org>

58. Mungall, C. J. *et al.* Use of model organism and disease databases to support matchmaking for human disease gene discovery. *Hum. Mutat.* **36,** 979–984 (2015).

59. Samwald, M., Chen, H., Ruttenberg, A. & Lim, E. Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience. *Artif. Intell.* (2009). at <http://linkinghub.elsevier.com/retrieve/pii/S0933365709001626>

60. Bota, M. & Swanson, L. W. BAMS neuroanatomical ontology: design and implementation. *Front. Neuroinform.* **2,** (2008).

61. Mihail Bota, L. W. S. Collating and Curating Neuroanatomical Nomenclatures: Principles and Use of the Brain Architecture Knowledge Management System (BAMS). *Front. Neuroinform.* **4,** (2010).

62. Larson, S., Fong, L., Gupta, A. & Condit, C. A formal ontology of subcellular neuroanatomy. *Frontiers in …* (2007). at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2525993/>

63. Magrane, M., Michele, M., Michele, M. & UniProt Consortium. UniProt Knowledgebase: a hub of integrated data. *Nature Precedings* (2010). doi:10.1038/npre.2010.5092.1

64. Roncaglia, P. *et al.* The Gene Ontology (GO) Cellular Component Ontology: integration with

SAO (Subcellular Anatomy Ontology) and other recent developments. *J. Biomed. Semantics* **4,** 20 (2013).

65. Halavi, M. *et al.* NeuroMorpho. Org implementation of digital neuroscience: dense coverage and integration with the NIF. *Neuroinformatics* **6,** 241–252 (2008).
66. Gupta, A., Bug, W., Marenco, L., Qian, X. & Condit, C. Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics* (2008). at <http://www.springerlink.com/index/rt2v4776456l1350.pdf>
67. Marenco, L., Wang, R., Shepherd, G. & Miller, P. The NIF DISCO Framework: Facilitating Automated Integration of Neuroscience Content on the Web. *Neuroinformatics* (2010). at <http://www.springerlink.com/index/C613M0L225P072G5.pdf>
68. Imam, F. T. *et al.* Development and use of Ontologies Inside the Neuroscience Information Framework: A Practical Approach. *Front. Genet.* **3,** 111 (2012).
69. Jeffrey, G. *et al.* SciCrunch: A cooperative and collaborative data and resource discovery platform for scientific communities. *Front. Neuroinform.* **8,** (2014).
70. Hall, D., Huerta, M. F., McAuliffe, M. J. & Farber, G. K. Sharing Heterogeneous Data: The National Database for Autism Research. *Neuroinformatics* **10,** 331–339 (2012).
71. Poldrack, R. A. *et al.* Discovering relations between mind, brain, and mental disorders using topic mapping. *PLoS Comput. Biol.* **8,** e1002707 (2012).
72. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8,** 665–670 (2011).
73. Zeng, Y., Wang, D. S., Zhang, T. L. & Xu, B. Frontiers | Linked Neuron Data (LND): A Platform for Integrating and Semantically Linking Neuroscience Data and Knowledge. (2014). at <http://www.frontiersin.org/10.3389/conf.fninf.2014.18.00017/event_abstract>

74. Begley, C. G. Six red flags for suspect work. *Nature* **497,** 433–434 (2013).

75. Peng, R. D. Reproducible Research in Computational Science. *Science* **334,** 1226–1227 (2011).

76. Friston, K. J. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2,** 189–210 (1994).

77. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23 Suppl 1,** S208–19 (2004).

78. Fischl, B. & Sereno, M. Cortical Surface-Based Analysis* 1:: II: Inflation, Flattening, and a Surface-Based Coordinate System. *Neuroimage* (1999). at <http://www.sciencedirect.com/science/article/pii/s1053-8119(98)90396-2>

79. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9,** 195–207 (1999).

80. Website. at <http://www.humanbrainmapping.org/files/2016/COBIDAS-ProposedFinal.pdf>

81. Wilbanks, J. T. Portable Legal Consent Overview. (2012).

82. Seneviratne, O. W. Augmenting the Web with Accountability. in *Proceedings of the 21st International Conference Companion on World Wide Web* 185–190 (ACM, 2012).

83. Choi, T. & Gouda, M. G. HTTPI: An HTTP with Integrity. in *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)* 1–6 (IEEE).

84. Benet, J. IPFS - Content Addressed, Versioned, P2P File System. *arXiv [cs.NI]* (2014). at <http://arxiv.org/abs/1407.3561>