Uma Plataforma Computacional para a Construção de Bancos de Dados para Experimentos de Neurociência*

Kelly Rosa Braghetto[1,2], Evandro Santos Rocha[1], Carlos Eduardo Ribas[1], Cassiano Reinert Novais dos Santos[1], Sueli dos Santos Rabaça[1], Margarita Ruiz Olazar[1]

[1]Centro de Pesquisa, Inovação e Difusão em Neuromatemática

[2]Departamento de Ciência da Computação - Instituto de Matemática e Estatística
Universidade de São Paulo – SP – Brazil

{kellyrb,erocha,ribas,cacorns,suelisr,mrolazar}@ime.usp.br

Resumo. *Dados científicos abertos são fundamentais para se ter uma ciência de melhor qualidade e de maior impacto. A criação de bancos de dados científicos abertos envolve vários desafios, como a criação de representações padronizadas para os dados e metadados de diferentes domínios do conhecimento e o desenvolvimento de recursos computacionais para auxiliar os cientistas na coleta e manutenção de dados de qualidade. Este artigo apresenta uma plataforma de software livre para o gerenciamento e compartilhamento de dados de experimentos em Neurociência. Essa plataforma permite registrar os dados e metadados de experimentos de forma segura e amigável, integrando registros de dados de diferentes tipos, como clínico, eletrofisiológico e comportamental.*

Abstract. *Open scientific data is fundamental to support better quality and higher impact reproducible science. The creation of open scientific databases involves a number of challenges, such as the creation of standardized representations of data and metadata from different domains of knowledge, as well as the development of computational resources to assist scientists in collecting and maintaining high-quality data. This paper presents a free software computational platform for the management and sharing of data from Neuroscience experiments. This platform allows to register data and metadata of experiments in a safe and user-friendly way, integrating data records of different types, such as clinical, electrophysiological and behavioral.*

1. Introdução

Com o aumento do uso de computação nas mais variadas áreas da ciência, os dados têm assumido um papel cada vez mais importante no processo de descoberta científica. Muitos dos resultados científicos que são divulgados hoje são embasados por dados digitais coletados ou gerados em experimentos científicos. Logo, é imprescindível que esses dados sejam confiáveis e estejam publicamente acessíveis, para que os resultados possam ser validados e reproduzidos. Apesar disso, em muitos domínios da ciência, como é o

*Esta plataforma foi desenvolvida como parte das atividades do Centro de Pesquisa, Inovação e Disseminação em Neuromatemática, financiado pela FAPESP (número do processo: 2013/07699-0). O trabalho também recebeu financiamento do CNPq (número do processo: 426579/2016-0).

caso da Neurociência, a disponibilização pública de dados de experimentos científicos ainda não é a regra, mas sim a exceção. A coleta de dados em experimentos é um trabalho difícil e dispendioso e ainda pouco reconhecido pela comunidade científica.

A representação e o armazenamento digital de dados científicos envolve diversos desafios. O projeto e a execução de um experimento científico inclui várias etapas, nas quais os seus parâmetros e a sua estrutura são definidos. No domínio da Neurociência, mais particularmente, caracterizar um experimento não é uma tarefa trivial. Existem diferentes tipos de experimentos (e.g., comportamentais, cognitivos, eletrofisiológicos e de neuroimagens), uma grande variabilidade na estrutura dos processos experimentais e uma alta heterogeneidade de formatos de dados coletados.

Para que um cientista possa fazer uma análise correta dos dados de um experimento em Neurociência ou seja capaz de reproduzi-lo, ele precisa conhecer informações sobre o processo experimental completo, ou seja, sobre como os dados foram coletados ou gerados. Além disso, há outras informações "ortogonais" ao processo experimental que também são indicadores importantes da qualidade dos dados coletados. Como exemplo, pode-se citar informações sobre o laboratório onde o experimento foi realizado, sobre os profissionais responsáveis pelo experimento e pelas coletas de dados e até mesmo publicações ou outros resultados decorrentes do experimento. Os dados sobre o processo experimental mais as informações ortogonais à realização de um experimento podem ser entendidos como *metadados* ou *dados de proveniência* dos dados experimentais.

1.1. Bancos de Dados na Neurociência

A construção, manutenção e curadoria de bancos de dados públicos em Neurociência é hoje considerada fundamental para um avanço mais efetivo na compreensão do funcionamento do cérebro e no tratamento de suas patologias. Esse movimento surgiu de maneira mais sistemática na área a partir da década de 90 [Chicurel 2000, Koslow 2000, Koslow 2002], quando se deu a primeira grande iniciativa de compartilhamento de dados coletados a partir de medidas de ressonância magnética funcional – o *International Consortium for Brain Mapping*[1]. Várias iniciativas de compartilhamento de dados de diferentes tipos têm sido implementadas desde então, principalmente em consórcios e grandes projetos como *Human Brain Project*[2], *International Neuroinformatics Coordinating Facility* (INCF)[3] e *Brain Research and Integrative Neuroscience Network* (BRAINnet)[4].

Apesar do crescente interesse na área, muitos dos bancos de dados de Neurociência disponíveis na atualidade possuem deficiências que dificultam o reúso dos dados e inviabilizam a aplicação de procedimentos computacionais para a descoberta automática de novos conhecimentos. Dentre essas deficiências, é possível destacar [Kötter 2001]: (i) dados de baixa qualidade, inconsistentes ou incompletos; (ii) bancos de dados que são "federações" de conjuntos heterogêneos de dados (com qualidades e estruturas diferentes e sem uma visão unificada dos dados); (iii) bancos de dados que são públicos mas não completamente abertos, ou seja, que impõem restrições ao acesso e ao reúso dos dados.

[1]International Consortium for Brain Mapping – http://www.loni.usc.edu/ICBM/
[2]Human Brain Project – https://www.humanbrainproject.eu
[3]International Neuroinformatics Coordinating Facility – https://www.incf.org/
[4]BRAINnet – http://www.brainnet.net/

1.2. Representação e Armazenamento de Dados em Neurociência

Muitos cientistas armazenam digitalmente os dados de seus experimentos como arquivos comuns, mantidos no sistema de arquivos de um computador. Os dados de proveniência, quando digitalizados, acabam virando arquivos texto (com dados não estruturados) ou planilhas sem uma estrutura padronizada. Essa forma de armazenamento dificulta a manutenção, a recuperação, o compartilhamento e o reúso dos dados.

Apesar da ausência de padrões, já existem iniciativas relacionadas à criação de diretrizes que definem quais são os dados que um pesquisador precisa reportar quando publica os resultados de um experimento em Neurociência. Exemplos de diretrizes desse tipo são a MINI (*Minimum Information about a Neuroscience Investigation*) [Gibson et al. 2009], a MINEMO (*Minimal Information for Neural Electromagnetic Ontologies*) [Frishkoff et al. 2011] e a *Guidelines for reporting an fMRI study* [Poldrack et al. 2008]. Essas *check lists* em geral apontam as informações que são consideradas importantes para a análise dos dados coletados e para a compreensão do experimento realizado. Entretanto, essas informações podem não ser suficientes para apoiar a reprodução do experimento ou o reúso dos seus dados.

Quanto à representação padronizada de dados de Neuroimagem, existem propostas tais como o *Neuroimaging Data Model* (NIDM)[5] e XCEDE-DM [Ghosh et al. 2012], modelos que capturam detalhes da aquisição e análise das imagens. Esses modelos de dados derivam do W3C PROV [Moreau et al. 2008], um modelo padrão para a troca de informações de proveniência. Para dados de eletrofisiologia, o modelo *Neurodata Without Borders* (NWB) [Teeters et al. 2015] se destaca. Olazar et al. [Ruiz-Olazar et al. 2016] caracterizaram e comparam algumas dessas iniciativas. Como nenhuma dessas propostas de padronização foi amplamente adotada na comunidade de Neurociência até agora, as ferramentas computacionais ainda são o principal recurso de suporte para cientistas interessados no intercâmbio de informações de Neurociência.

1.3. Nova Plataforma para a Construção de Bancos de Dados de Experimentos

Para que se tenha bancos de dados de experimentos que atendam requisitos mínimos de qualidade, é preciso que os cientistas tenham à sua disposição ferramentas computacionais que os amparem nas suas tarefas rotineiras de condução dos experimentos e de coleta e análise dos dados relacionados. Apesar de já existirem repositórios públicos para o armazenamento e compartilhamento de dados na área de Neurociência, ainda há uma forte carência por ferramentas de software que mantenham, de forma integrada e categorizada, todos os dados coletados em um experimento e a suas informações de proveniência.

Com o objetivo de propor uma solução para esse problema, este artigo apresenta uma plataforma de software de apoio à criação e à disponibilização pública de bancos de dados de experimentos em Neurociência. Essa plataforma é composta por duas ferramentas de software – o *Neuroscience Experiments System* (NES) e o *NeuroMat Open Database* (NeuroMat DB). Essas ferramentas são uma iniciativa do Centro de Pesquisa e Difusão em Neuromatemática (CEPID NeuroMat), financiado pela FAPESP.

O NES é um software livre para o gerenciamento de dados de experimentos de Neurociência. O NES garante que todo dado de experimento registrado no banco de

[5]Neuroimaging Data Model – http://nidm.nidash.org/

dados mantido por ele esteja devidamente acompanhado de suas informações de proveniência, impondo um formato comum para a representação e armazenamento dos dados de experimentos de um mesmo laboratório ou grupo de pesquisa. Com isso, os dados armazenados têm melhor qualidade e ficam mais fáceis de serem compartilhados publicamente e reusados. O NES permite que dados anonimizados de um experimento sejam enviados para o NeuroMat DB, que disponibiliza dados publicamente por meio de um portal Web.

2. Experimentos em Neurociência

Experimentos em Neurociência podem ser de diferentes tipos, como, por exemplo, comportamentais, de eletrofisiologia ou de neuroimagem. Experimentos de eletrofisiologia geralmente envolvem Eletroencefalografia (EEG), Estimulação Magnética Transcraniana (EMT) ou Eletromiografia (EMG), enquanto que os de neuroimagens costumam gerar imagens por Ressonância Magnética (RMI) ou Ressonância Magnética Funcional (fMRI).

Em um experimento em Neurociência, os sujeitos (humanos ou animais) geralmente são separados em grupos. Cada grupo pode ser submetido a um conjunto de condições experimentais específicas. Cada tipo de experimento envolve uma preparação específica para sua realização, como, por exemplo, configurações no equipamento de coleta de sinais eletrofisiológicos e a colocação de eletrodos em locais previamente definidos do corpo dos sujeitos do experimento. Todas as definições sobre um experimento, incluindo definições sobre os objetivos, a descrição dos grupos de sujeitos que serão testados, as condições experimentais às quais os grupos serão submetidos e os tipos de coletas de dados que serão realizados, são chamadas pelos cientistas de *protocolo experimental*.

O processo experimental geralmente é composto por três etapas: (i) *planejamento do experimento*, (ii) *coleta e armazenamento de dados* e (iii) *análise dos dados*. Na fase de *planejamento do experimento*, é realizada a preparação para realização do experimento, onde devem ser identificados os tipos de dados que serão coletados e a definição do protocolo experimental. Depois, um grupo de sujeitos é selecionado para participar do experimento e a coleta dos dados começa a ser realizada. Utilizando o protocolo experimental como guia, na fase de *coleta e armazenamento de dados* é realizado o registro dos dados primários do experimento para cada participante. Dados primários são os que estão em sua forma bruta (i.e., como foram recebidos das suas fontes). A coleta de dados primários pode ser realizada de várias formas, como, por exemplo: o preenchimento de questionários projetados para o experimento; o uso de equipamentos para a captura de dados, como EEG e MRI; observação do comportamento do participante em resposta às condições experimentais às quais ele é exposto. Na última fase, que é a de *análise de dados*, os dados primários são processados e analisados e podem gerar novos dados, chamados de *dados derivados*. Dados derivados e seus processos de geração também devem ser armazenados e documentados para serem posteriormente reusados ou reproduzidos.

3. O *Neuroscience Experiments System* (NES)

O *Neuroscience Experiments System* (NES) é uma ferramenta de software livre para auxiliar neurocientistas no gerenciamento dos dados de seus experimentos. O NES coleta, estrutura e organiza os dados de todas as etapas do processo experimental de um experimento em Neurociência. A Figura 1 ilustra como o NES se integra ao processo experimental. Na fase do desenho experimental, o NES permite registrar os metadados do experimento, como a descrição do protocolo experimental, a configuração dos equipamentos

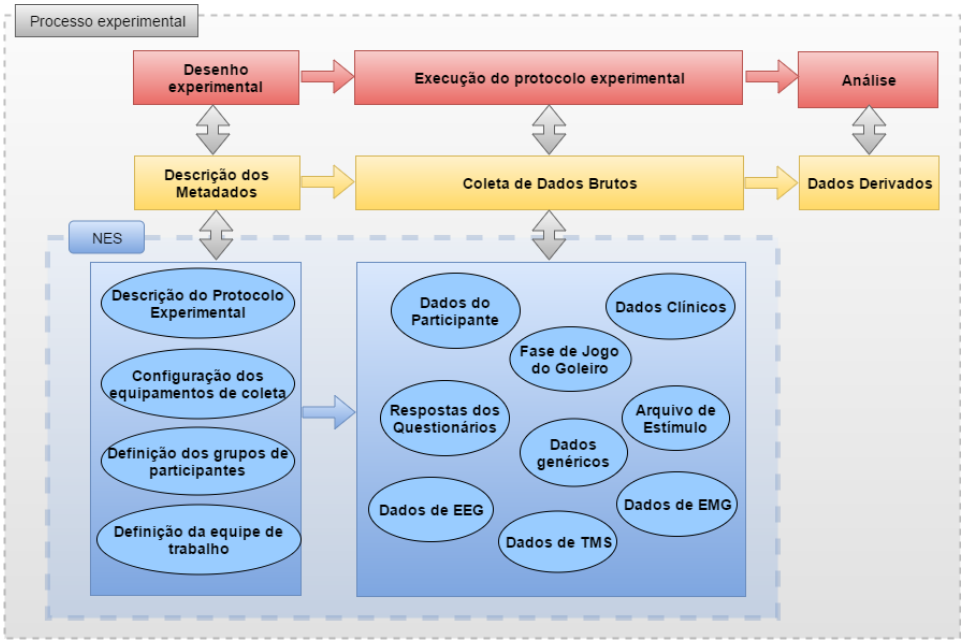

Figura 1. Diagrama da integração do NES no processo experimental

de coleta dos dados, a definição dos grupos de participantes e da equipe de trabalho, entre outras informações que descrevem o contexto do experimento.

Para caracterizar dados experimentais e metadados de vários tipos de experimentos, o NES foi desenvolvido para contemplar a coleta de dados de diversas naturezas, como dados de EEG, EMG, EMT, dados de estímulos (i.e., auditivos, visuais, etc.), dados genéricos (i.e., estabilometria, cinemática, comportamentais, etc.), dados de respostas de questionários, dados vindos da integração com o Jogo do Goleiro[6], dados sobre os participantes (sujeitos) dos experimentos, assim como outras informações que podem ser inseridas de forma textual ou como arquivos anexados ao experimento. Assim, o NES fornece uma plataforma estruturada, robusta e abrangente, com suporte a dados de proveniência, para possibilitar a rastreabilidade dos dados e a reprodução dos experimentos.

O modelo de dados do NES está alinhado com diretrizes para reportar experimentos em eletrofisiologia, como as já citadas MINI e MINEMO. O NES também trabalha com vários formatos de arquivos usados pela comunidade de Neurociência. Em particular, é interoperável com o formato *Neuroscience Without Border* (NWB), uma das iniciativas mais promissoras para a representação padronizada de dados de eletrofisiologia. O modelo de armazenamento usado no NES permite que dados provenientes de fontes de informação variadas sejam convertidos em formatos interoperáveis (e.g., CSV, JSON, PDF, etc.) que podem ser facilmente exportados para uso em diferentes plataformas computacionais.

Por ter sido desenvolvido como um sistema Web, o NES possui uma interface amigável e pode ser executado em vários tipos de dispositivos, como computadores *desktop*, *tablets*, *smartphones*, etc., com a mesma qualidade de apresentação, provendo maior fa-

[6]O Jogo do Goleiro é um jogo desenvolvido pelo CEPID NeuroMat para estudar a forma como o cérebro humano reconhece padrões: `http://game.numec.prp.usp.br/`

cilidade de uso. Os dados gerenciados pelo NES são armazenados em um banco de dados relacional, usando o *PostgreSQL*, de forma a garantir facilidade de manutenção e segurança dos dados. O armazenamento dos dados coletados por sistemas externos é feito por meio de arquivos nos formatos padronizados existentes. Os dados de proveniência desses últimos também são armazenados, facilitando a recuperação posterior.

3.1. Funcionalidades do NES

Experimentos em Neurociência envolvem uma grande heterogeneidade de formatos de dados e metadados complexos. Para esses requisitos, o NES fornece as funcionalidades:

Registro de Participantes: para ajudar no controle dos participantes, o NES mantém o cadastro de informações pessoais, dados sociodemográficos, história social e avaliações médicas. Às avaliações médicas, é possível associar diagnósticos com o uso do Código Internacional de Doenças (CID) e anexar possíveis exames realizados.

Gerenciamento de Questionários: o NES está integrado o LimeSurvey[7] para gerenciar a administração de questionários eletrônicos usados na coleta de dados em experimentos. Com o uso de questionários eletrônicos, outras dados mais específicos ou avaliações longitudinais podem ser coletadas de forma estruturada.

Gerenciamento de Experimento: essa funcionalidade envolve o registro e a configuração do experimento, assim como a descrição do protocolo experimental. Um protocolo experimental é descrito no NES como um *workflow* não-automatizado. Dessa forma, o NES consegue representar todas as condições experimentais como passos de um processo, onde cada passo pode ser algo como: uma instrução para o participante, uma aplicação de questionário, a apresentação de um estímulo, uma tarefa para o participante, uma tarefa para o experimentador, um conjunto de passos, etc. Um conjunto de passos é uma forma de organizar sub-passos, que podem ser realizados de forma paralela ou sequencial. Uma vez que o protocolo experimental tenha sido criado, é possível realizar coletas de dados referentes aos participantes do experimento. O NES é capaz de lidar com dados eletrofisiológicos (e.g., EEG e EMG) coletados em vários formatos usados pela comunidade de Neurociência. No NES, cada dado coletado em um experimento está sempre associado ao sujeito do qual foi coletado e a uma etapa específica do protocolo experimental.

Exportação de Dados: o NES permite exportar todos os dados e metadados dos experimentos que ele armazena. A exportação inclui os dados dos sujeitos do experimento (i.e., respostas dos questionários, dados clínicos, dados primários, etc.) e metadados sobre o protocolo experimental (i.e., descrição do experimento e das etapas do protocolo, configuração de equipamentos e anotações feitas por cientistas). Além disso, é possível realizar filtros pelos dados dos participantes, como sexo, diagnóstico e idade. NES exporta os dados textuais e numéricos organizados em arquivos em formatos textuais puros (e.g., CSV e JSON). Dados de EEG podem ser exportados no formato NWB.

3.2. Integração com o NeuroMat Open Database (NeuroMat DB)

O NES é um software que deve ser instalado em um servidor Web para gerenciar e manter os dados de experimentos de um laboratório ou grupo de pesquisa em Neurociência em particular. Para manter a segurança dos dados, o NES usa criptografia de dados e mecanismos de controle de acesso baseado em perfis de usuários.

[7]LimeSurvey – https://www.limesurvey.org/

Para compartilhar dados e metadados de experimentos de forma pública, o CEPID NeuroMat disponibiliza o *Neuromat Open Database* (NeuroMat DB). O portal Web do NeuroMat DB[8] é uma plataforma de acesso aberto para compartilhamento e busca de dados e metadados de experimentos de Neurociência. Através do NES, um pesquisador pode enviar dados e metadados de seus experimentos para o serviço Web do NeuroMat DB. Os dados dos participantes são anonimizados antes de serem enviados; nenhum dado sensível é enviado do NES para o banco de dados aberto. Quando um novo conjunto de dados de um experimento chega ao Neuromat DB, um comitê de curadoria analisa se os dados são apropriados para publicação. Os conjunto de dados aprovados pelo comitê são disponibilizados publicamente no portal Web do Neuromat DB.

3.3. Licença de Uso e Outros Softwares Livres

Tanto o NES quanto o NeuroMat DB têm uma arquitetura completamente aberta e foram desenvolvidos a partir de ferramentas de software livre, como a linguagem de programação Python, o arcabouço Web Django, o arcabouço de *front-end* Bootstrap e o sistema gerenciador de banco de dados PostgreSQL. A licença do NES e do Neuromat DB é a Mozilla Public License versão 2.0[9]), que dá total liberdade para uso e alterações dos softwares. O código fonte, a documentação e o status do desenvolvimento do NES e do NeuroMat DB podem ser vistos nos seguintes endereços, respectivamente: `https://github.com/neuromat/nes` e `https://github.com/neuromat/portal`.

4. O Uso do NES na Construção de Bancos de Dados

O Laboratório de Pesquisa em Neurociências e Reabilitação (LNR) do Instituto de Neurologia Deolindo Couto (INDC) da UFRJ, em colaboração com o CEPID NeuroMat, está trabalhando na investigação dos mecanismos de plasticidade cerebral e na avaliação de preditores da resposta ao tratamento de reabilitação em pacientes com lesões do plexo braquial – um conjunto de nervos que inervam os membros superiores. O LNR está usando o NES para armazenar e documentar os dados coletados em seus estudos, que são constituídos principalmente por registros eletrofisiológicos, respostas a questionários clínicos e dados comportamentais. Uma versão parcial do conjunto de dados desses estudos já está disponível publicamente no portal Web do NeuroMat DB.

O NES também está sendo usado no gerenciamento de dados dos estudos conduzidos na Rede AMPARO[10], uma iniciativa do CEPID NeuroMat para promover a melhora na qualidade de vida de pessoas vivendo com Doença de Parkinson no Brasil e de seus familiares. Esses estudos envolvem, principalmente, dados clínicos, respostas de questionários e dados comportamentais capturados com o Jogo do Goleiro.

5. Considerações Finais

Existem vários desafios relacionados à criação de bancos de dados abertos na área de Neurociência, como a ausência de padrões para a representação de dados de experimentos, consequência da complexidade e variabilidade da estrutura dos protocolos experimentais. Mas para que se possa disponibilizar publicamente dados experimentais, é preciso garantir que eles sejam registrados de forma estruturada, com consistência e completude,

[8]NeuroMat DB – `http://neuromatdb.numec.prp.usp.br/`
[9]Mozilla Public License versão 2.0 – `https://www.mozilla.org/en-US/MPL/2.0/`
[10]Rede AMPARO – `https://amparo.numec.prp.usp.br/`

conjuntamente com suas informações de proveniência. Isso é essencial para que dados desse tipo possam ser entendidos e reusados.

Nesse contexto, o *Neuroscience Experiments System* (NES) e o *NeuroMat Open Database* (NeuroMat DB) são importantes contribuições para a comunidade de Neurociência, por constituírem um plataforma de software livre amigável e, ao mesmo tempo, poderosa para o registro e compartilhamento de dados experimentais e suas informações fundamentais de proveniência. Essas ferramentas já estão sendo usadas na criação de bancos de dados que visam amparar progressos na compreensão do funcionamento cerebral e no diagnóstico e tratamento de doenças neurológicas.

Atualmente, estamos estendendo as duas ferramentas a fim de habilitá-las a gerenciar dados derivados (ou seja, dados resultantes de um processamento computacional aplicado sobre os dados primários). Também desejamos incorporar às ferramentas a capacidade de registrar protocolos e coletas de experimentos envolvendo neuroimagens.

Referências

[Chicurel 2000] Chicurel, M. (2000). Databasing the brain. *Nature*, 406(6798):822–825.

[Frishkoff et al. 2011] Frishkoff, G., Sydes, J., Mueller, K., Frank, R., Curran, T., Connolly, J., Kilborn, K., Molfese, D., Perfetti, C., and Malony, A. (2011). Minimal information for neural electromagnetic ontologies (MINEMO): A standards-compliant method for analysis and integration of event-related potentials (ERP) data. *Standards in Genomic Sciences*, 5(2):211–223.

[Ghosh et al. 2012] Ghosh, S., Nichols, N., Gadde, S., Steffener, J., and Keator, D. (2012). Xcede-dm: A neuroimaging extension to the W3C provenance data model. In *Front. Neuroinform. Conference Abstract: 5th INCF Congress of Neuroinformatics*.

[Gibson et al. 2009] Gibson, F., Overton, P. G., Smulders, T. V., Schultz, S. R., Eglen, S. J., Ingram, D., Panzeri, S., Bream, P., Sernagor, E., Cunningham, M., Echtermeyer, C., Simonotto, J., Kaiser, M., Swan, D. C., and Lord, P. (2009). Minimum information about a neuroscience investigation (MINI): electrophysiology. *Nature Precedings*.

[Koslow 2000] Koslow, S. H. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience*, 3:863–866.

[Koslow 2002] Koslow, S. H. (2002). Sharing primary data: a threat or asset to discovery? *Nature Reviews Neuroscience*, 3(4):311–313.

[Kötter 2001] Kötter, R. (2001). Neuroscience databases: tools for exploring brain structure–function relationships. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1412):1111–1120.

[Moreau et al. 2008] Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V., and Varga, L. (2008). The provenance of electronic data. *Communications of the ACM*, 51(4):52–58.

[Poldrack et al. 2008] Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., and Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *Neuroimage*, 40(2):409–414.

[Ruiz-Olazar et al. 2016] Ruiz-Olazar, M., Rocha, E. S., Rabaça, S. S., Ribas, C. E., Nascimento, A. S., and Braghetto, K. R. (2016). A review of guidelines and models for representation of provenance information from neuroscience experiments. In *International Provenance and Annotation Workshop*, pages 222–225. Springer.

[Teeters et al. 2015] Teeters, J. L., Godfrey, K., Young, R., Dang, C., Friedsam, C., Wark, B., Asari, H., Peron, S., Li, N., Peyrache, A., et al. (2015). Neurodata without borders: Creating a common data format for neurophysiology. *Neuron*, 88(4):629–634.