

Perspectives Workshop: Digital Scholarship and Open Science in Psychology and the Behavioral Sciences

Edited by

Alexander Garcia Castro¹, Janna Hastings², Robert Stevens³, and Erich Weichselgartner⁴

1 Technical University of Madrid, ES, alexgarcia@gmail.com

2 European Bioinformatics Institute – Cambridge, GB, hastings@ebi.ac.uk

3 University of Manchester, GB, robert.stevens@manchester.ac.uk

4 ZPID – Trier, DE, wga@zpid.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15302 “Perspectives Workshop: Digital Scholarship and Open Science in Psychology and the Behavioral Sciences”. This workshop addressed the problem of facilitating the construction of an integrative digital scholarship and open science infrastructure in psychology and the behavioral sciences by utilizing the Web as an integrative platform for e-Science. A particular focus was on sharing research data and experiments to improve reproducibility. The participants presented first steps in this direction in their communities, and worked out an initial action plan for establishing digital scholarship and open science more broadly.

Seminar 19–24. July, 2015 – <http://www.dagstuhl.de/15302>

1998 ACM Subject Classification J.4 Social and Behavioral Sciences, I.7.4 Electronic Publishing, H.3.5 Online Information Services, H.3.7 Digital Libraries

Keywords and phrases Digital Scholarship, Open Science, Psychology, Behavioral Sciences, e-Science

Digital Object Identifier 10.4230/DagRep.1.1.1

Edited in cooperation with Christoph Lange

1 Executive Summary

Garcia Castro, Alexander; Hastings, Janna; Lange, Christoph; Stevens, Robert; Weichselgartner, Erich

License © Creative Commons BY 3.0 Unported license

© Garcia Castro, Alexander; Hastings, Janna; Lange, Christoph; Stevens, Robert; Weichselgartner, Erich

Researchers across many domains have invested significant resources to improve transparency, reproducibility, discoverability and, in general, the ability to share and empower the community. Digital Scholarship and Open Science are umbrella terms for the movement to make scientific research, data and dissemination accessible to all members of an inquiring society, amateur or professional. Digital infrastructures are an essential prerequisite for such open science and digital scholarship; the biomedical domain illustrates this symbiosis. An impressive digital infrastructure has been built; this allows us to correlate information from genomes to diseases, and, by doing so, to support personalized medicine. A high degree of interdisciplinary work was necessary in building this infrastructure; the large quantities of data being produced, the high degree of interrelatedness, and, most of all, the need for this



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Perspectives Workshop: Digital Scholarship and Open Science in Psychology and the Behavioral Sciences, *Dagstuhl Reports*, Vol. 1, Issue 1, pp. 1–24

Editors: Alexander Garcia Castro, Janna Hastings, Robert Stevens, and Erich Weichselgartner



Dagstuhl Reports

DAGSTUHL
REPORTS

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

mingling of many types of data in a variety of forms forged this collaboration across the community and beyond.

The Behavioral Sciences, comprising psychology but also psychobiology, criminology, cognitive science and neuroscience, are also producing data at significant rates; by the same token, understanding mental health disorders requires correlating information from diverse sources – e.g. cross-referencing clinical, psychological, and genotypic sources. For example, flagship projects such as the Brain Activity Map (BAM, also known as the BRAIN initiative¹) are generating massive amounts of data with potential benefit to mental health and psychology; conversely, projects like BAM could benefit from information currently being generated by psychologists. Our ability to make continued progress in understanding the mind and brain depends on finding new ways to organize and synthesize an ever-expanding body of information.

The *‘Digital Scholarship and Open Science in Psychology and the Behavioral Sciences’* Dagstuhl Perspectives Workshop was conceived with one problem in mind: that of facilitating the construction of an integrative infrastructure in Psychology and Behavioral Sciences. The motivation for this workshop was to *‘foster the discussion around the problem of understanding the Web as an integrative platform, and how e-science can help us to do better research.’* With these points in mind, we gathered an interdisciplinary group of experts, including computer scientists, psychologists and behavioral scientists. In their research, they are addressing issues in data standards, e-science, ontologies, knowledge management, text mining, scholarly communication, semantic web, cognitive sciences, neurosciences, and psychology. Throughout the Workshop, this group worked on devising a roadmap for building such an interoperability layer.

The seminar started with a number of keynote sessions from well-known authorities in each area to introduce the necessary background and form a common baseline for later discussions. A core theme that emerged was the cross-domain challenge in establishing a common language. We jointly undertook the effort to define an integrative scenario illustrating how digital infrastructures could help psychologists and behavioral scientists to do research that takes advantage of the new digital research landscape. In order to achieve this, the computational scientists needed to better understand the current working practices of the psychologists. For instance, the nature and structure of their data and experiments; moreover, computer scientists needed to understand the flow of information, from the conception of an idea, through defining a study plan, executing it and finally having the investigation published. They learned that the work of psychologists and behavioral scientists strongly relies on questionnaires and experiments as ways of collecting data, and on statistics as a tool for analyzing data, and that the replicability of experiments is a key concern. In a similar vein, psychologists and behavioral scientists needed concrete examples illustrating how computer science enables FAIR (= findable, accessible, interoperable and reusable) infrastructures that allow researchers to discover and share knowledge – bearing in mind data protection issues.

Two break-out groups were organized. The purpose was to have a full picture of digital scholarship in action when applied to psychology and behavioral investigations, most importantly e-science assisting researchers in sharing, discovering, planning and running investigations. The full research life cycle had to be considered. Both groups worked up their respective scenarios independently. The visions were then exchanged in an inter-group meeting. Interestingly, various issues arose when discussing the specifics from each vision for digital scholarship; for instance, the importance of understanding scholarly communication

¹ <http://www.braininitiative.nih.gov/>

beyond the simple act of getting one's results published. Furthermore, the need to integrate tools into platforms where researchers could openly register their projects and plan and manage their workflows, data, code and research objects, was extensively discussed. Within this framework, the need for controlled vocabularies, standards for publishing and documenting data and metadata, persistent identifiers for datasets, research objects, documents, organizations, concepts and people, open APIs to existing services and instruments, and reporting structures were understood; these elements were articulated in the examples where the researchers and research were at the center of the system. Discussions also addressed fears in the community and thus the need to open up the current research landscape in small steps.

The seminar proved to be a fertile discussion field for interdisciplinary collaborations and research projects across previously disparate fields with the potential of significant impact in both areas. The need for a digital infrastructure in psychology and behavioral sciences was accepted by all the attendants; communicating this message with a clear implementation vision to funding agencies, professional societies and the community in general was identified as a key priority. It was decided that we needed another meeting in 2016; during that follow-up, the emphasis should be on developing a research agenda. As this is a relatively new topic in psychology and behavioral sciences, it was also decided to contact publishers and professional organizations, e.g. the Sloan Foundation, the APA and the APS, and work with them in conveying the message about increasing openness. If we want to understand how cognition is related to the genome, proteome and the dynamics of the brain, then interoperability, data standards and digital scholarship have to become a common purpose for this community. Funding has to be made available, initially for an assessment of the uptake of existing key resources and infrastructures, and then for implementing further Digital Scholarship and Open Science infrastructures as well as for building the skills in a community that is not yet widely familiar with the relevant enabling technologies. Finally, once sufficient technical support is in place, sustainable incentives for sharing research objects should be put in practice.

2 Table of Contents

Executive Summary

<i>Garcia Castro, Alexander; Hastings, Janna; Lange, Christoph; Stevens, Robert; Weichselgartner, Erich</i>	1
---	---

Overview of Talks

Mental illnesses, knowledge representation and data sharing <i>Xavier Aime</i>	6
The Human Behaviour Project <i>Dietrich Albert</i>	6
Data Archiving and Sharing Confidential Data <i>George Alter</i>	7
#dsos requires a digital infrastructure <i>Bjoern Brembs</i>	8
VIVO – Connect, Share, Discover. An open source, semantic web software system and ontology for representing scholarly work <i>Mike Conlon</i>	8
Research Objects for improved sharing and reproducibility in Psychology and Behavioural Sciences <i>Oscar Corcho</i>	9
Estimating the Reproducibility of Psychological Science <i>Susann Fiedler</i>	9
Goals of the Seminar on Digital Scholarship and Open Science in Psychology and the Behavioral Sciences <i>Alexander Garcia Castro</i>	10
‘Don’t Publish, Release’ Revisited <i>Paul Groth</i>	10
Open Science Lessons Learned at Mendeley <i>William Gunn</i>	11
The role of standards and ontologies in tackling reproducibility <i>Janna Hastings</i>	11
Developing reproducible and reusable methods through research software engineering <i>Caroline Jay</i>	12
Open science, mega analyses and problems in understanding the genetics of psychiatric disorders <i>Iris-Tatjana Kolassa</i>	13
Advancing Psychology and Behavioral Sciences in Brazil and World Wide <i>Silvia Koller</i>	14
Scholarly Communication and Semantic Publishing: Technical Challenges, and Recent Applications to Social Sciences <i>Christoph Lange</i>	15

Defining the Scholarly Commons: Are We There Yet: Summary of my presentation
and some thoughts on the workshop
Maryann Martone 19

Cognitive ontologies, data sharing, and reproducibility
Russell Poldrack 20

Open Journal Systems: Introduction, Preview, and Community
Alec Smecher 21

Open Science and Digital Scholarship at Elsevier
Daniel Staemmler 21

Open data and the need for ontologies
Robert Stevens 22

Infrastructural Services for the Scientific Community provided by the American
Psychological Association
Gary VandenBos 23

Hijacking ORCID
Hal Warren 23

3 Overview of Talks

3.1 Mental illnesses, knowledge representation and data sharing

Xavier Aime (LIMICS INSERM U 1142 – Paris, FR)

License © Creative Commons BY 3.0 Unported license
© Xavier Aime

Joint work of Aimé, Xavier; Richard, Marion; Charlet, Jean; Krebs, Marie-Odile
Main reference Richard, M.; Aimé, X.; Krebs, M. & Charlet, J. Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts *Studies in Health Technology and Informatics*, 2015, 210, 221-223

Mental illnesses have a major impact on public health but their pathophysiology remains largely unknown and their treatments insufficient. One of the main difficulties is that these disorders are defined only on the basis of clinical syndromes issued from historical descriptions, now with consensual criteria in the international classifications (CIM 10 from WHO or DSM-IV from American Psychiatric Association). An additional hallmark of psychiatry is the apparent lack of specificity of the biological markers or risk factors, when identified, and the overlap of symptoms across diagnostic categories. There is thus an urgent need to be able to define more precisely the phenotype, or profile of anomalies, at the individual level, by taking into account numerous and heterogeneous ways of characterization, collected in large clinical databases. The domain of psychological disorders raises several challenges that need to be addressed, as large amount and diversity of sources and nature of information, the evolutivity of symptoms and diachronic trajectories of mental disorders across a person's life span, and special requirements of all human rights documents and protection of privacy. Research in this area is data intensive, which means that data sets are large and highly heterogeneous; therefore the use of inappropriate models would lead to inappropriate (if not flawed) results. Clinical data is complex, non trivial, and redundant. To create knowledge from such data, researchers must integrate and share these large and diverse data sets. This presents daunting computer science challenges such as representation of data that is suitable for computational inference (knowledge representation with an ontology such as OntoPsychia [1]), and linking heterogeneous data sets (data integration – unfortunately, data integration and sharing are hampered by legitimate and widespread privacy concerns). The use of an ontology, associated with dedicated tools, will allow also (1) to perform semantic research in Patient Discharges Summary (PDS), (2) to represent the comorbidity, (3) to index PDS for the constitution of cohorts, and (4) to identify resistant patient's profiles.

References

- 1 M. Richard and X. Aimé and M.O. Krebs and J. Charlet. Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts. *Studies in Health Technology and Informatics*, Vol. 210, pp. 221–223, 2015.

3.2 The Human Behaviour Project

Dietrich Albert (TU Graz, AT)

License © Creative Commons BY 3.0 Unported license
© Dietrich Albert

Understanding, predicting and modifying human behaviour of individuals and groups in artificial and natural environments belong to the greatest challenges facing 21st century

basic & applied sciences and research & development (R&D). Rising to these challenges, we can (1) gain profound insights into human behaviour and its underlying structures in all aspects, develop new methods for behavioural diagnosing and predicting as well as new treatments for behavioural changes and preventions, and (2) build revolutionary new computing technologies. For the first time, modern information and communication technologies (ICT) enable of tackling these goals. Thus, the project aims to achieve a multi-modal, integrated understanding of behavioural structures and functioning through the development and use of ICT.

By bringing together

- machine readable theoretical and empirical content,
- modern wireless sensors and manipulanda technology,
- big behavioural data (offline and online),
- techniques for dynamic representations of contexts and environments,
- open and adaptive database methodology,
- social media approaches,
- technologies for machine learning and big data analytics,
- semantic technologies etc.

totally new technologies for

- scalable, process-oriented, complex real time simulations

will be developed in

- strong co-operation of the behavioural sciences and the computer sciences.

Inherent

- process-oriented evaluative,
- ethical components, and
- gender aspects

will be implemented.

These technologies will realise simulations of individual and group behaviour in different contexts and environments, large-scale collaboration and data sharing, federated analysis of behavioural data, and the development of complex integrated computing systems. Through the projects ICT platforms, scientists, stakeholders, and engineers will be able to perform diverse experiments and share knowledge with a common goal of unlocking human behaviour. With an unprecedented cross-disciplinary scope, the project will integrate and stimulate behavioural science, computing, and social science, will unify theory and practice, and benefit the global scientific community dealing with humans. The development and use of ICT will pave the way for the project's ultimate goal, the simulation of human behaviour in terms of both individuals and groups in artificial and real settings.

3.3 Data Archiving and Sharing Confidential Data

George Alter (University of Michigan – Ann Arbor, US)

License © Creative Commons BY 3.0 Unported license
© George Alter

My presentation outlined the certification of “trusted digital repositories” and a framework for sharing confidential data. Trusted digital repositories are expected to make data discoverable, meaningful, usable, trustworthy, and persistent. This means that repositories must have procedures to document and preserve data and policies to sustain their institutional viability.

“Deductive disclosure” refers to re-identifying subjects from a combination of their characteristics in a data set. Data repositories use a range of measures that providing access to the research community while minimize the risk of disclosure. Procedures for sharing confidential data can be characterized under the headings: safe data (anonymization), safe places (data enclaves), safe people (legal agreements), and safe outputs (vetting computed results). Since these measures are intrusive and hinder researchers, the severity of the measures should be weighed against the disclosure risks for each data set.

3.4 #dsos requires a digital infrastructure

Bjoern Brembs (Universität Regensburg, DE)

License © Creative Commons BY 3.0 Unported license
© Bjoern Brembs

Main reference Brembs B, Button K and Munafò M (2013) Deep impact: unintended consequences of journal rank. *Front. Hum. Neurosci.* 7:291. doi: 10.3389/fnhum.2013.00291

URL <http://journal.frontiersin.org/article/10.3389/fnhum.2013.00291/full>

Access is only one of many functionalities that are badly broken in our scientific infrastructure. Our literature would lose little of its functionality if we carved it in stone, took pictures of it and put them online. Our data – if it is made accessible at all – all too often rests in financially insecure databases. And our scientific code is hardly available at all, with no institutional infrastructure to speak of. If the vision of digital scholarship that is open by default is to become a reality, we need to raise funds to build the digital infrastructure supporting digital open scholarship. On the local level, we have developed proofs-of-concept, demonstrating the time-saving potential of such an infrastructure. On the international institutional level, I argue that we need to use the funds currently wasted on subscriptions to implement this infrastructure as soon as possible.

3.5 VIVO – Connect, Share, Discover. An open source, semantic web software system and ontology for representing scholarly work

Mike Conlon (University of Florida, US)


License © Creative Commons BY 3.0 Unported license
© Mike Conlon

draft by
Christoph;
get Mike's
approval

VIVO is an open source, semantic web application, ontology and community providing standard data and tools for representing scholarship and using data about scholarship. VIVO integrates and reuses institutional data about an organization, its scholars, its grants and projects, its publications and scholarly works, as well as its teaching and engagement. From this it can generate reports, portfolios and curricula vitae, as well as visualizations. The knowledge aggregated in a VIVO installation can help to find experts, to analyze networks, and to answer ad hoc queries. It is reusable because it is exported as 5-star Linked Open Data. VIVO has been widely adopted by research institutions around the world, predominantly in the US.

3.6 Research Objects for improved sharing and reproducibility in Psychology and Behavioural Sciences

Oscar Corcho (Technical University of Madrid, ES)

License  Creative Commons BY 3.0 Unported license

© Oscar Corcho

URL <http://www.slideshare.net/ocorcho/research-objects-for-improved-sharing-and-reproducibility>

When a researcher is working on a specific experiment, no matter what his/her scientific discipline is, a large amount of entities are used during the research process. This includes papers that have been read, input datasets, scripts, pieces of code, spreadsheets, output data, etc. Some time later, when this researcher goes back to all this material to resume this work, or another researcher wants to make use of it for another piece of research, he/she will normally find it very difficult to find all the material that was used at the time of the original investigation, to understand the purpose of some of those scripts, etc.

Research Objects have been proposed in the literature as a mechanism to aggregate all that material, making it more easily discoverable, providing identifiers to all these elements, and including metadata to understand better all these elements. More information available at <http://www.researchobject.org/> and details of the Research Object Model at [1]

During this Dagstuhl meeting we have had the opportunity to understand the type of resources that should be included in the most common types of Research Objects in the areas of Psychology and Behavioural Sciences, so as to propose in the future a Research Object profile that can be used in this area.

References

- 1 Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina Hettne, Raul Palma, Eleni Mina, Oscar Corcho, José Manuel Gómez-Pérez, Sean Bechhofer, Graham Klyne, Carole Goble. Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*. Vol. 32, pp. 16–42, 2015. doi:10.1016/j.websem.2015.01.003.

3.7 Estimating the Reproducibility of Psychological Science

Susann Fiedler (MPG – Bonn, DE)

License  Creative Commons BY 3.0 Unported license

© Susann Fiedler

Joint work of Open Science Collaboration


Main reference Open Science Collaboration, Estimating the Reproducibility of Psychological Science, Science, in press

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects ($M_r = .197$, $SD = .257$) were half the magnitude of original effects ($M_r = .403$, $SD = .188$), representing a substantial decline. Ninety-seven percent of original studies had significant results ($p < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, combining original and replication results left 68% with significant effects. Correlational tests suggest that replication success was

better predicted by the strength of original evidence than by characteristics of the original and replication teams.

3.8 Goals of the Seminar on Digital Scholarship and Open Science in Psychology and the Behavioral Sciences

Alexander Garcia Castro (Technical University of Madrid, ES)

License  Creative Commons BY 3.0 Unported license
© Alexander Garcia Castro


draft by
Christoph; get
Alex' approval

The “Digital Scholarship and Open Science in Psychology and Behavioral Sciences” seminar has two specific goals, namely:

- To foster and initiate the discussion about open science and digital scholarship in psychology and the behavioral sciences, addressing specific issues such as data standards, interoperability, knowledge representation, ontologies, linked data
- To identify useful experiences from other domains, requirements, issues to be addressed. More importantly, to define a common vision, a road map for this community to build cyber infrastructures in support of open science and digital scholarship.

3.9 ‘Don’t Publish, Release’ Revisited

Paul Groth (Elsevier Labs – Amsterdam, NL)

License  Creative Commons BY 3.0 Unported license
© Paul Groth

At the 2013 Beyond the PDF 2 conference, Prof. Carole Goble’s presented the notion that research communication should be more like software development. It is now evident, that many of the components in that vision are now a reality. We can iterate, test, debug, execute and build upon our science using similar tools. For example, journals such as Cognition allow for the pre-registration of materials. Experiments performed primarily in-silico can be tracked and reproduced using virtual machines. Version systems such as GitHub can be used to keep track of versions, histories and dependencies. Such systems allow for forking, starring, branching, which can be used to derive credit. Data repositories allow for experimental data to be stored and cited. Notebook environments such as Jupyter enable even creative analysis to be shared and reproduced. As we go forward, these components will become seamlessly interconnected. Such interconnection will enable new transparency and visibility of the process of science. This increased visibility requires science to develop new norms, namely, a new stance of constructive criticism. The openness enabled by these technologies demands that we recognize that there are bugs and they can be fixed. We should embrace the inherent interaction of science.

3.10 Open Science Lessons Learned at Mendeley

William Gunn (Mendeley Ltd. – London, GB)

License © Creative Commons BY 3.0 Unported license
© William Gunn

Open Science is a new word for a old practice. What we now call Open Science used to just be called science. In the early days, science wasn't funded by national agencies, of course, but there were societies of learned gentlemen who used to meet to share their results, write letters back and forth, etc. How and why did that change? As technology grew in importance to society, science professionalized and the discussion of science also had to professionalize. Scholarly societies began to turn to companies like Elsevier for their ability to run a journal, manage the operations, coordinate the peer review, publishing, distribution, and so on. It made a lot of sense, back when publishing to a worldwide audience necessarily had to be a difficult and expensive endeavor, to let companies derive profit in exchange for the hard work of editing, producing, and distributing research reports. Then the internet came along. We're not entirely sure what scientific publishing on the Web will look like in 20 years, but we're pretty sure that it won't continue to look like it has for the past 100+ years.

At Mendeley we have learned some lessons about Open Science, and we are continuously drawing from other successful examples of leveraging the Web. One way to make publication of research on the Web more like publication of other things on the Web is to make it open, indexed, shareable, and available in multiple forms. Because the research article is not the work, it's the report of the work. It is essentially an advertisement that you have done a certain amount of work, but that which is contained in your publication is only a very lossy compression of your work, and your work consists of data generated, software created, methods developed, and the broader impacts on society you've had. The obvious thing to do is to connect these articles, these reports of the work, to the work itself. Another example of an innovation to come by leveraging the Web for open science is dynamic views of the primary data, instead of just the 2D representation that was generated by the authors at the point of publication. Detailed protocols for generating the data, software and virtual environments that render the data into the view chosen by the author or into another view, facilitate reader understanding of the strengths and limitations of the data as collected and improve reproducibility. At Mendeley, we will continue to innovate in these ways through working on the Resource Identification Initiative, the Reproducibility Initiative, and building integrations with ELN tools like Hivebench so that the provenance of an experiment is captured along with the final outcome, allowing the work to be placed in context and built upon.

3.11 The role of standards and ontologies in tackling reproducibility

Janna Hastings (European Bioinformatics Institute – Cambridge, GB)


License © Creative Commons BY 3.0 Unported license
© Janna Hastings

is a multi-faceted challenge. Standards and ontologies play an important role in many of these facets. Reproducibility is enhanced through the provision of raw data in open access repositories such that the analyses leading to results can be entirely reproduced by different researchers and the data can be reused for different research questions. However, for raw

data to be truly reusable, it must be presented in an accessible format and annotated in a standardised fashion using shared ontologies. A minimum amount of information about the way in which the data was generated needs to be provided, as well as important contextual information about the entities that were investigated and the purpose of the study. One of the hardest problems in achieving the wide exchange and sharing of well-annotated data is the sociological challenge of bringing communities together in order to create, and proliferate the use of, good standards. A standard is no use unless it is widely adopted by the full community. Ensuring adoption and usage requires the development of good tools supporting such usage and the tireless promotion of standards compliance across the full community.

3.12 Developing reproducible and reusable methods through research software engineering

Caroline Jay (University of Manchester, GB)

License  Creative Commons BY 3.0 Unported license

© Caroline Jay

Joint work of Jay, Caroline; Haines, Robert

Discussions around open science and digital scholarship often focus on the important topics of creating and applying data standards, and achieving a robust infrastructure to support research. That scientists will follow standard, or at least well-defined, methods and operating procedures is a given – a crucial first step in ensuring research is reproducible.

In reality, research methods in psychology are often far from standard; they continually and necessarily evolve to meet the challenges of understanding new forms of behaviour and interaction. In the domain of human-computer interaction (HCI), this is particularly true, as traditional paradigms for investigating behaviour often cannot be directly applied to technology use.

In many psychological studies, software is firmly embedded in both the data collection and analysis processes: packages such as E-Prime and Tobii Studio are popular tools for ensuring that reaction time and gaze data measurements are taken reliably, and are straightforward to interpret. Both these software tools are proprietary, however, and whilst this results in stability that is helpful from the perspective of reproducibility, it is less useful from the perspective of open science.

Truly achieving reproducibility is hard. The authors have been striving to ensure their science is open for several years, but issues such as incomplete raw data, data that cannot be published for ethical reasons, the use of proprietary software, hard-to-decipher analysis scripts and unavailable experimental materials have all proved barriers to reaching this goal (see for example, [1]).

At the University of Manchester, and in particular as part of the EPSRC-funded IDInter-action project (EP/M017133/1), we are trying to address these challenges, by developing open-source software methods that not only make it easy to reproduce experimental results, but are also suitable for reuse and extension. Underlying our new approach is one crucial factor: the recognition of software engineering as a first class citizen in the research process. By paying attention to the usability and sustainability of software during the experimental design process, rather than treating it as an afterthought (or ignoring it completely), we hope to develop tools and methods that can be used to demonstrate the reproducibility of our own work, and support further experiments in the future.

Convincing others of the utility of ‘research software engineering’, and embedding it in the mainstream of scientific activity, is likely to require a significant cultural shift. Both scientists and research funders must recognise that the additional resources necessary to support this activity are vital to the future of science. Excellence in software engineering practice is essential to developing reproducible and reusable methods; scientists (for now, at least), are unlikely to be able to achieve this alone. As such, people with a focus on software development are as vital to producing genuinely reproducible computational research as people with a focus on the science itself.

References

- 1 C. Jay, A. Brown, S. Harper. *Predicting whether users view dynamic content on the world wide web*. *ACM TOCHI*, 20(2), 2013.

3.13 Open science, mega analyses and problems in understanding the genetics of psychiatric disorders

Iris-Tatjana Kolassa (Universität Ulm, DE)

License  Creative Commons BY 3.0 Unported license
© Iris-Tatjana Kolassa

For a better understanding of the etiology of psychiatric disorders and in order to develop new medication and successful treatments we need to combine data originating from both clinical psychology and genetics, and combine studies from around the world to increase sample sizes. An infrastructure that allows easy data sharing and exchange of knowledge will be highly beneficial for this purpose. So-called ‘mega analyses’ combine participant-level data from multiple different original studies to reach sample sizes of up to tens of thousands of subjects (in contrast to traditional ‘meta analyses’, which combine summary results and parameter estimates on aggregate levels). First such mega analyses have already been conducted; however, their results have been disappointing so far. A recent mega analysis of the Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium concluded that even a sample of 18.759 independent and unrelated subjects with and without major depressive disorder is still underpowered to detect genetic effects typical for complex traits. Besides sample size, one reason why mega-analyses have failed might be that they do not consider gene x environment interactions, which are crucial if one wants to understand psychiatric diseases. One example which demonstrates this particularly well is posttraumatic stress disorder (PTSD) – a disorder that requires experiencing a traumatic event and thus is an example of an inherent gene x environment interaction. In experiencing a traumatic event, a fear network is built up that contains sensations, emotions, cognitions and interoceptive experiences associated with the traumatic situation. With increasing number of traumatic event types experienced, the fear network increases in size, and specific triggers can reactivate multiple traumatic events experienced. With increasing traumatic load, the lifetime prevalence of PTSD reaches 100%, the symptom severity increases and the probability of spontaneous remission decreases. However, genetic factors interact with trauma: Some studies suggest that in the case of low trauma load, genetic factors might play an important role, while in the case of extremely high traumatic load, the environmental factor is more influential than the genetic constitution of an individual. One important problem of all current studies on the role of genes in the etiology, symptomatology and treatment of PTSD is that it is not easy to quantify the environmental factor traumatic load,

in particular across various populations and studies. However, initial evidence shows that it needs to be considered not only in the etiology of PTSD, but also in studies assessing the dependency of treatment effects on genetic factors. Furthermore, the gene x environment equation needs to be broadened, e.g. epigenetic modifications need to be considered when assessing the effects of genotypes, and genetic pathway analyses might be more helpful than single candidate gene association studies. Open access genetic data would be helpful for combining data of various studies, for reanalyzing previous studies given new knowledge gained, and finally to spot mistakes in statistical analyses by the scientific community, as the statistics of gene x environment interactions is complex, and frequently errors occur in how factors that might influence the dependent variable are controlled for, e.g., if the groups differ significantly in this covariate (see [1]). Misunderstanding analysis of covariance, *Journal of Abnormal Psychology*). However, while there are many reasons why open data would be highly beneficial for this field of research, the protection of individual data is a particularly sensitive topic when studying traumatic event types experienced in highly sensitive populations (e.g. victims of wars, genocide or other atrocities) as well as when analyzing not only single nucleotide polymorphisms but also whole genome and epigenome data.

References

- 1 Miller, G.A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110, 40–48.

3.14 Advancing Psychology and Behavioral Sciences in Brazil and World Wide

Silvia Koller (Federal University of Rio Grande do Sul, BR)

License  Creative Commons BY 3.0 Unported license
© Silvia Koller

Main reference Full Professor and Chair of the Center for Psychological Studies of At-Risk Populations in the Department of Psychology at the UFRGS. Scientific Chair of Brazilian Virtual Library in Psychology

URL <https://scholar.google.com.br/citations?user=KF11ZRkAAAAJ&hl=pt-BR&oi=ao>

The seminar “Digital Scholarship and Open Science in Psychology” and Behavioural Sciences took place in the week of 20th to July 25th 2015.

The multi and inter disciplinary workshop was attended by scientists of psychology, behavior analysis, computer, biologists and biomedical sciences. Principles such as accessibility, sharing, interoperability and the possibility of multiple uses of the knowledge produced in several areas were the basis for discussions during the whole week. It was emphasized issues related to ontology of information systems, open access knowledge and data sharing in science.

The breakthrough perspective of knowledge is intrinsically linked to the possibility of opening and systematic and ongoing sharing of related objects to research, beyond the one that just occur when of the publication of scientific articles.

There was continuing emphasis on the need for fluidity and systematic opening of knowledge generated and tested, so that the science of behavior and psychology effectively may produce and improve the quality of life of human beings through knowledge. To get to such a possibility is needed clear description, systematic, standardized and rigorous terms, methods and procedures, beyond just data and results.

In Brazil, SciELO and BVS Psychology are examples of broad dissemination of science in these areas. New perspectives, such as the creation of the Science Data Repository DadoPsi – <http://dadopsi.bvs-psi.org.br>, further enable the advancement of Psychology and Behavioral Sciences in Brazil.

Moreover, it is very well received and accepted among scientists participating in the seminar, the fact that Brazil favors open access to the content of their magazines through tools, such as Scielo and Pepsic.

All kinds of open science, either through new open tools, new forms of standardization of terms and methods based on well organized ontologies and metadata, will contribute to the advancement of the area.

3.15 Scholarly Communication and Semantic Publishing: Technical Challenges, and Recent Applications to Social Sciences

Christoph Lange (Universität Bonn, DE)

License  Creative Commons BY 3.0 Unported license
© Christoph Lange

This contribution presents an overview on current *technical* challenges to digital scholarship and open science (DSOS), and to visions for overcoming them the near future. With a focus on technology, this overview is largely domain-independent; however, it gives some specific insight into the domain of social science.

Overview

The overarching goal of my DSOS research is to enable scholars to share knowledge in a FAIR (findable, accessible, interoperable, reusable²) way. The key assumption underlying my research agenda is that FAIR sharing of scholarly knowledge is possible with information and communication technology that supports the complete process of research and scholarly communication without “media disruptions”³ between its individual steps. My proposed solution is to employ *linked data* technology for preserving information created by scientists in experiments, by authors while writing, and by reviewers while commenting on a paper, in an explicit way to enable intelligent services to act on it – and to provide data-driven services to readers, authors and reviewers inside the familiar environment of a paper.

In the following, I point out media disruptions between the tools that so far support the scientific process. I argue why linked data has the potential to overcome these disruptions. I present first results and the further agendas of our ongoing projects in this field, covering the perspective of collaborative work environments as well as the publication of (meta)data that enable services. Finally, this extended abstract makes the case for combining both strands of research to support the idea of open access, which should not only be seen from a legal perspective, with a sustainable technical foundation.

² This notion of FAIRness has originally been introduced for research data by the FORCE11 initiative on the Future of Research Communication and e-Scholarship; see their guiding principles at <https://www.force11.org/node/6062>.

³ The term “media disruptions” is my free translation of the German term “Medienbruch”, which refers to a point in information processing where the carrier medium of the information changes. This change typically results in loss of information, dropping information quality, or at least inefficiency.

Problem Statement

The increasingly collaborative scientific process, from a project plan to the design of an experiment, to collecting data, to interpreting them and writing down that interpretation in a paper, to submitting that paper for peer review, to publishing an accepted paper, to, finally, its consumption by readers who find, read and cite it, is insufficiently supported by contemporary information systems. They support every *individual* step, but media disruptions between steps cause inefficiency or even loss of information. Examples include:

- Word processors lack direct access to data.
- There is no assistant that would automatically recommend authors where to submit their paper (i.e. to a high-profile event whose topic the paper matches and where the paper has a realistic chance to be accepted).
- Reviewers do not provide feedback in the same environment in which authors will be revising their papers.
- Open access web publishing is restricted to document formats designed for paper printing but neglecting the Web's accessibility and interactivity potential.
- Readers, seeing a single, frozen view of the underlying data in a paper, are unable to access the full extent and the further dimensions of the data.
- Information that helps to assess the quality of a scientific publication, such as the peer reviews it received, the history of the venue (conference or journal) in which it has been published, and information on the context in which it has been cited, are scattered over different places, or not even available in a machine-comprehensible format.

The Potential of Linked Data Technology

My research is based on the assumption that web technology, in particular *semantic* web and linked data technology, can address these problems, for two reasons:

1. its potential to integrate heterogeneous systems and heterogeneous data: Isolated solutions, such as tools for publishing data on the Web for easy retrieval and visualisation, exist in preliminary manifestations in the social sciences and other domains, but have not been integrated into tools for writing, reviewing and publishing articles.
2. its approach of making the structure and semantics of data and documents explicit to machines: document browsers that use articles as interactive interfaces to related information on the Web, tools that make knowledge FAIR and even remixable, as well as tools that assist writers in making their texts machine-comprehensible with little additional effort, have been deployed successfully in the life sciences and other fields.

My research aims at transferring these ideas to the social sciences and beyond by integrating existing data and publication management services into a web-based collaborative writing environment that publishers can set up to support all types of end users throughout the publication process: authors, reviewers and readers. To ensure acceptance by decision makers and end users, specifically non-technical users, and to take advantage of existing solutions, even if isolated, new collaboration environments should be compatible with the existing solutions. If a seamless integration is not feasible, compatibility should at least be established by import/export interfaces or by well-defined migration paths. Such flexible levels of integration are easiest to achieve based on free, open, well-documented and extensible systems that already do part of the job. For example, the collaborative document editor

Fidus Writer⁴ and the Open Journal Systems submission and review management system⁵ provide a stable technical foundation for such integration efforts.

Ongoing Efforts and First Results

We are working on a *collaborative writing environment* as outlined above in the concrete setting of the ‘Opening Scholarly Communication in the Social Sciences’ (OSCOSS) project, which will run from autumn 2015 to autumn 2017 and involves, besides the University of Bonn, the GESIS social science institute⁶ as an application partner. In this project, we aim at securing user acceptance primarily by respecting the characteristics of the traditional processes social scientists are used to: web publications must have the same high-quality layout as print publications, and information must remain citable by stable page numbers. To ensure we meet these requirements, we will work closely with the publishers of ‘methods, data, analyses’ (mda) and ‘Historical Social Research’ (HSR), two international peer reviewed open access journals published by GESIS, and build early demonstrators for usability evaluation. The OSCOSS system will initially provide readers, authors and reviewers with an alternative, thus having the potential to gain wider acceptance and gradually replace the old, incoherent publication process of the participating journals.

Secondly, *data and metadata* of which scientists could take advantage, while doing their research and writing about it, is increasingly available on the Web; however, there are two key limitations:

1. The datasets provide insufficient details and are often merely superficially machine-comprehensible because valuable information is lost before or during their publication: for example, ...
 - there are dataset registries, such as da|ra for the social sciences⁷, that make datasets retrievable and citable by publishing metadata about them, but so far they do not effectively enable the maintainers of datasets to publish the *content* of their datasets in a machine-comprehensible and thus FAIR way.
 - Also, publication databases hardly allow for assessing the excellence of a publication, researcher or venue in a way more comprehensive than counting citations, as further contextual information is not published (e.g., acceptance rates) or not yet easy to exploit (e.g., information on the structure and dynamics of research communities or on the context in which sources are cited).
2. Comprehensive services such as a conference/journal recommender assistant would have to utilise data and metadata from multiple sources; however, there is so far little interlinking between such data sources.

Our work in the context of the OpenAIRE2020 European project (OpenAIRE = Open Access Infrastructure for Research in Europe⁸) running from 2015 to 2018, where the University of Bonn is leading the Linked Open Data (LOD) activities, on the so far two editions of the Semantic Publishing Challenge⁹, and my work as technical editor of the CEUR-WS.org open access publication service for computer science addresses these problems. In OpenAIRE2020,

⁴ <http://www.fiduswriter.org>

⁵ <https://pkp.sfu.ca/ojs/>

⁶ <http://www.gesis.org>

⁷ Registration agency for social and economic data; see <http://www.da-ra.de>

⁸ <http://www.openaire.eu>

⁹ <https://github.com/ceurws/lod/wiki/SemPub2015>

we are concerned with publishing metadata about all EU-funded research projects, their results (publications and datasets, soon also software) and their participating organisations and persons as LOD¹⁰, and in a second step to interlink them with related datasets, or to enrich them with information from other open datasets with which they cannot be interlinked. The 2014 and 2015 Semantic Publishing Challenges have addressed the problem of extracting information from publications and proceedings that would help to better assess their quality, e.g., information on the history of event series and on citation contexts. These challenges work on the data of CEUR-WS.org, paving the path towards publishing them as LOD. The other part of my work at CEUR-WS.org is, similarly to the work planned in the OSCOSS project mentioned above, concerned with reducing the loss of information in the publication process by avoiding media disruptions (e.g., by enabling direct generation of high-quality proceedings volumes from the EasyChair submission system used widely in computer science¹¹), and with lowering the barrier to publishing proceedings and papers in a machine-comprehensible and thus FAIR way for authors and chairs who are not really “non-technical” (as we are in computer science), but, as experience shows, busy or lazy and therefore not willing to waste time.

An Integrated Technical Foundation for Open Access

In summary, both strands of research outlined above – integrated collaboration environments for readers, authors, reviewers, and the provision of higher-quality research data and metadata – are expected to yield promising results, partly supported by project funding until 2017/2018. However, their *synthesis* calls for a new project, which promises the following two benefits:

1. data-driven services for readers, authors and reviewers, all accessible from inside the familiar environment of a paper without loss of information caused by media disruptions and without loss of efficiency caused by switching tasks. Such services include:
 - reusing and remixing data while reading/writing/reviewing. Making the links between tables and figures and their underlying datasets explicit enables interactive document players to support users in exploring different scenarios beyond the restricted scope chosen by the author.
 - recommendations of citations based on the local context of the author’s current position in the document, and recommendations of publication venues based on the structure and full text of a paper.
2. an advanced collaboration environment that generates, during the normal flow of interacting with it, and at no extra cost, machine-comprehensible metadata that give others – open access repository maintainers as well as immediate readers – FAIR access to the scientific results produced inside the environment.

Same as “open data” in the narrow sense is merely a legal framework for making data reusable, while *linked* data technology enables its practical realisation¹², the tight integration of research data and metadata with collaborative writing and reviewing environments will serve as a technical companion to the legal concept of *open access*. It will make journals more “open” (in terms of FAIRness) that are, legally, open access already, and it has the potential to serve as an incentive for turning “closed” journals into open access ones.


¹⁰ <http://lod.openaire.eu>

¹¹ See the ceur-make tool at <https://github.com/ceurws/ceur-make>

¹² This is practically explained at <http://5stardata.info>.

3.16 Defining the Scholarly Commons: Are We There Yet: Summary of my presentation and some thoughts on the workshop

Maryann Martone (UC – San Diego, US)

License  Creative Commons BY 3.0 Unported license
© Maryann Martone

In this presentation, I went through experiences in the neurosciences with aggregating and searching across large amounts of data, based on our experiences in designing and operating the Neuroscience Information Framework (NIF). I particularly focused on the importance of ontologies for providing a conceptual backbone for search and organization of data across many different scales and disciplines. Because there is no single data type or technique that defines neuroscience, without the conceptual underpinnings, there is not way to bring together the different types of information, nor for searching across the hundreds of millions of records contained in thousands of databases and data sets.

The Neuroscience Information Framework, and its sister project SciCrunch, also provide a practical data set for examining the current resource landscape. NIF has been cataloging and tracking research resources (data, tools, materials) for over 8 years. We see that funders are very willing to set up these resources, but a remarkable number of them grow stale or disappear because of lack of support.

Although in any endeavor, it is common for a large number of initiatives to be started and only some to take off, given the funding difficulties currently facing biomedicine, this ‘launch and languish’ model is not very cost effective. I talked a bit out SciCrunch, our configurable data portal technology, which allows communities to create their own portals, customized to their needs and branded with their own identity. However, SciCrunch portals are connected on the back end by a shared data infrastructure, so that any data added or improved propagates throughout the network automatically.

I also discussed the power of web-based annotation as a means to add a connecting and interactive knowledge layer on top of our scholarly output. Hypothesis is a non-profit that has developed the capability of annotating the web. Anyone by installing a plug in can highlight text on a web page or PDF and add an annotation. It is an open tool being engineered for an emerging W3C standard for web annotation.

With Hypothesis, we can open up new information channels across static publications, and add critical and currently missing knowledge about things like reproducibility. Alec Smecher of Open Journal Systems showed how Hypothesis was integrated into the OJS platform.

I concluded by sharing some practical lessons that we’ve learned about open science. One of the most important is that people are important to this endeavor: data sharing and open science doesn’t just magically happen. It needs champions and the societies need to support it. Data resources become interesting when there is a lot of data, so means to match requirements for data sharing to our current incentive system is key.

Finally, I believe that the period of letting a thousand flowers bloom in creating these resources is past. We have to invest in existing infrastructure, e.g., institutional repositories and community repositories, to make them better. Our current trend is to look at them, find fault with them, and then start again. But this practice leads to ‘partially built cars’. What do I mean by that? Think of current research infrastructures as cars. If our current funding model built fully functional cars, then some would win and some would lose but we’d still have cars to drive. If our current system didn’t build cars but car parts, then someone could take these interoperable pieces and build a car that works. But what we currently do is build

partially built, non-interoperable cars. So we may be able to limp along in one or two, but none can fully thrive.

We think that it is time for the community to start to come together around the idea of the Scholarly Commons—that is, the Set of protocols, principles, best practices, API's and standards that govern flow of scholarly research object. The ultimate goal is to make research objects (the sum total of research output) FAIR: Findable, accessible, interoperable, reusable. Through organizations like FORCE11 (Future of Research Communications and e-Scholarship), we are getting closer to be able to articulate what is required for research communities to become part of the commons.

Thoughts: This workshop was an extremely valuable discussion forum for bringing several of the concepts in my talk into clearer focus. The exercise where we designed a future system and then realistically assessed where we stood led to the concept of making communities 'e Science ready' as opposed to overpraising what we can do today. What became clearer, and what I have used in talks since then is that there are some things that communities need to make the transition from non-digital to digital. Because these researchers in our current reward system are not likely to accrue significant benefits in the beginnings, the 'asks' cannot be overly onerous.

So what is required:

1. People, concepts, instruments and materials need to enter the eScience world with a persistent identifier attached. Efforts like ORCID and the Resource Identification Initiative are making headway and should be supported. If a community doesn't have an open ontology or controlled vocabulary, they need to support the creation of one. If they can't make their instruments, e.g., questionnaires, eScience enabled (i.e., unique ID, network accessible), then they need appropriate tools to do so. The latter tools should help their science by making it easier for them to create what they need and not hinder it.
2. Data needs to be made potentially accessible and recoverable. The strongest argument for data sharing right now is the transparency argument, that is, all data that was produced in the course of the study needs to be made available and potentially recoverable in the future. That means having the data hosted in an appropriate repository and having at least minimal standardized metadata. We have the means to do both and it doesn't take a lot of the researchers time to work with curators to achieve this. Proper norms about what should be shared publicly (or not) and when need to be developed in conjunction with the community, but if data are properly deposited and stewarded, then the data can be shared when these agreements are reached. If we don't make the data potentially recoverable, then we lose it for all time.

3.17 Cognitive ontologies, data sharing, and reproducibility

Russell Poldrack (Stanford University, US)

License  Creative Commons BY 3.0 Unported license
© Russell Poldrack

In my talk, I outline the need for formal ontologies to describe cognitive processes, and provide an overview of the Cognitive Atlas project, which aims to develop such an ontology. I describe the structure of the Cognitive Atlas, focusing particularly on the different classes of entities (tasks and concepts) that are represented in the knowledge base. The Cognitive

Atlas has been used to annotate the OpenfMRI database of neuroimaging data, and this annotation has been used to support ontology-driven data mining. I also outline ongoing work in our group on reproducibility in the context of neuroimaging, discussing the threats to reproducibility that are inherent in current practices and describing the work of the Stanford Center for Reproducible Neuroscience, which is developing a new resource to allow researchers to better quantify the reproducibility of their findings.

3.18 Open Journal Systems: Introduction, Preview, and Community

Alec Smecher (Simon Fraser University – Burnaby, CA)

License  Creative Commons BY 3.0 Unported license
© Alec Smecher

Open Journal Systems is a widely used open source web application providing a complete journal publishing workflow, emphasizing (but not exclusive to) Open Access publishing by automating the time-consuming and expensive workflow process. It is written and maintained by the Public Knowledge Project (PKP), <http://pkp.sfu.ca>, with contributions of code, translations, etc. from a diverse community of contributors.

In the 13 years since OJS 1.0 was released, it has grown from a proof of concept into a mature piece of infrastructure helping to facilitate the publishing of many thousands of journals in over 30 languages.

More recently, PKP has successfully introduced a new platform for managing scholarly monographs and edited volumes. Open Monograph Press (OMP) was also used as an opportunity to pioneer a rewrite of aspects of OJS that were aging or in need of updating to keep pace with new scholarly publishing trends.

In August 2015 PKP will unveil a beta release of OJS 3.0, including numerous technical and workflow improvements.

At the 2015 Dagstuhl conference on Digital Scholarship and Open Science in Psychology and the Behavioral Sciences, crossover discussions between open science and open source have frequently arisen, both in terms of the importance of Free and Open Source Software to scientific replicability, and of the potential for tools and workflows from the open source software development community to be studied and potentially introduced into the future practices of psychology research.

3.19 Open Science and Digital Scholarship at Elsevier

Daniel Staemmler (Elsevier Publishing – Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Daniel Staemmler

Elsevier is concerned about fostering open science and digital scholarship in its numerous journals. Examples include:

- *Cognition* applies a data sharing policy requiring every author to either provide a link to a data repository, or to submit supplementary data with their article.
- The *Journal of Research in Personality* fosters transparency in the scientific process by requiring “all authors [to] abide by standard principles of research conduct, including making data available to other researchers”.

draft by
Christoph;
get Daniel's
approval

- The implementation of the TOP guidelines (Transparency and Openness Promotion) is currently under evaluation.
- *SoftwareX* is a new journal for getting software index, discoverable, accessible and cited.
- *MethodsX* is a new journal to get method customization indexed, discoverable, accessible and cited.

3.20 Open data and the need for ontologies

Robert Stevens (University of Manchester, GB)

License  Creative Commons BY 3.0 Unported license
© Robert Stevens

This is an abstract for “Digital Scholarship and Open Science in Psychology and the Behavioural Sciences”, a Dagstuhl Perspectives Workshop (15302) held in the week commencing 20 July 2015. The workshop brought together computer scientists, computational biologists and people from the behavioural sciences. The workshop explored eScience, data, data standards and ontologies in psychology and other behavioural sciences. This abstract gives my view on the advent of eScience in parts of biology and the role open data and metadata supplied by ontologies played in this change.

There is a path that can be traced with the use of open data in the biological domain and the rise in the use of ontologies for describing those data. Biology has had open repositories for its nucleic acid and protein sequence data and controlled vocabularies were used to describe those data. These sequence data are core, ground truth in biology; all else comes from nucleic acids and, these days, the environment. As whole genome sequences became available, different organism communities found that the common vocabulary used to represent sequences facilitated their comparison at that level, but a lack of a common vocabulary for what was known about those sequences blocked the comparison of the knowledge of those sequences. Thus we could tell that sequence A and sequence B were very similar, but finding that the function, processes in which they were involved and where they were to be found etc. was much more difficult, especially for computers. Thus biologists created common vocabularies, delivered by ontologies, for describing the knowledge held about sequences. This has spread too many types of data and many types of biological phenomenon, from genotype to phenotype and beyond, so that there is now a rich, common language for describing what we know about biological entities of many types.

At roughly the same time was the advent of eScience. The availability of data and tools open and available via the Web, together with sufficient network infra-structure to use them, led to systems that co-ordinated distributed resources to achieve some scientific goal, often in the form of workflows. Open tools, open data, open standards, open, common metadata all contribute to this working, but it can be done in stages; not all has to be perfect for something to happen – just availability of data will help, irrespective of its metadata. Open data will, however provoke the advent of common data and metadata standards, as people wish to do more and do it more easily.

In summary, we can use the FAIR principles (Findable, Accessible, Interoperable and re-usable) to chart this story. First we need data and tools to be accessible and this means openness. Metadata, via ontologies, also have a role to play in this accessibility – do we know what those data are etc.? Metadata has an obvious role in making tools and data findable – calling the same things by the same term and knowing what those terms mean makes things

findable. The same argument works for interoperable tools and data.

3.21 Infrastructural Services for the Scientific Community provided by the American Psychological Association

Gary VandenBos (American Psychological Association, US)

License © Creative Commons BY 3.0 Unported license
© Gary VandenBos

My input to this workshop is based on my experience as the Publisher of the American Psychological Association¹³ in Washington, DC, USA, and as the co-Editor of the Archives of Scientific Psychology¹⁴, an open methods, collaborative data sharing, open access journal. I have designed electronic knowledge dissemination products for the field of psychology since 1984, including moving the Psychological Abstracts from a print product to a CD-based electronic product to PsycINFO¹⁵, a streaming Internet product. I also developed PsycARTICLES¹⁶ (a full-text journal article database), PsycBOOKS¹⁷ (a full-text book and book chapter database), PsycTESTS¹⁸ (a measurement instrument database), and PsycTHERAPY¹⁹ (a streaming video database of psychotherapy demonstrations). I am the Editor of the Publication Manual of the American Psychological Association²⁰. I have been an advocate for data sharing since 1990, and have served on many governmental and association task forces on data sharing – including the recent TOP Guidelines developed by the Center for Open Science²¹.

3.22 Hijacking ORCID

Hal Warren (Vedatek Knowledge Systems, US)

License © Creative Commons BY 3.0 Unported license
© Hal Warren

The Open Researcher and Contributor ID (ORCID) is a subset of an International Standard Name Identifier (ISNI), a 16 digit number that serves as a persistent identifier. This persistence turns human data into individually known machine readable data that can remain until the end of our civilization. The Internet was first the domain of scholars. ORCID was created as a means to disambiguate works of scholarly authors. It is time to broaden the audience for ORCID to everyone, taking advantage of persistence to join all our public personas into a single identifier, ORCID. By using the ORCID record to connect my Uniform Resource Identifiers (URIs) such as my Facebook account, my Twitter account as well as all of my email addresses, each instance of me can serve as a legitimate identifier of me which

¹³ <http://www.apa.org/>

¹⁴ <http://www.apa.org/pubs/journals/arc/>

¹⁵ <http://www.apa.org/pubs/databases/psycinfo/>

¹⁶ <http://www.apa.org/pubs/databases/psycarticles/>

¹⁷ <http://www.apa.org/pubs/databases/psycbooks/>

¹⁸ <http://www.apa.org/pubs/databases/psyc-tests/>

¹⁹ <http://www.apa.org/pubs/databases/psyc-therapy/>

²⁰ <http://www.apastyle.org/manual/>

²¹ <https://cos.io/top/>

can be verified against the ORCID record. ORCID adds credibility and provenance to whom I am online by joining different silos of my data so that machines can better reason on it.

Scholarly publishers are positioned to take advantage of ORCID by adding advanced machine reasoning to better structure disambiguated data. By assisting authors with the ORCID update process, needed infrastructure to support Research Object-based academic credit will emerge. My annotations are automatically connected to me regardless of the channel in which they are created. I become more complete.

By joining our health, financial, contribution and consumption data through ORCID, we create a trusted digital corpus with new capacity. Vedatek Knowledge Systems is hijacking ORCID for ordinary citizens, to improve their quality of life through the use of new sensor data and to augment the growth of local community connections.