



Volume 5, Issue 7, July 2015

Algorithms and Scheduling Techniques to Manage Resilience and Power Consumption in Distributed Systems (Dagstuhl Seminar 15281) <i>Henri Casanova, Ewa Deelman, Yves Robert, and Uwe Schwiegelshohn</i>	1
The Constraint Satisfaction Problem: Complexity and Approximability (Dagstuhl Seminar 15301) <i>Andrei A. Bulatov, Venkatesan Guruswami, Andrei Krokhin, and Dániel Marx</i> ...	22
Digital Scholarship and Open Science in Psychology and the Behavioral Sciences (Dagstuhl Perspectives Workshop 15302) <i>Alexander Garcia Castro, Janna Hastings, Robert Stevens, and Erich Weichselgartner</i>	42

ISSN 2192-5283

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <http://www.dagstuhl.de/dagpub/2192-5283>

Publication date

January, 2016

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 3.0 DE license (CC BY 3.0 DE).



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Aims and Scope

The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.

In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Editorial Board

- Bernd Becker
- Stephan Diehl
- Hans Hagen
- Hannes Hartenstein
- Oliver Kohlbacher
- Stephan Merz
- Bernhard Mitschang
- Bernhard Nebel
- Bernt Schiele
- Nicole Schweikardt
- Raimund Seidel (*Editor-in-Chief*)
- Arjen P. de Vries
- Michael Waidner
- Reinhard Wilhelm

Editorial Office

Marc Herbstritt (*Managing Editor*)
Jutka Gasiórowski (*Editorial Assistance*)
Thomas Schillo (*Technical Assistance*)

Contact

Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
<http://www.dagstuhl.de/dagrep>

Digital Object Identifier: 10.4230/DagRep.5.7.i

Algorithms and Scheduling Techniques to Manage Resilience and Power Consumption in Distributed Systems

Edited by

Henri Casanova¹, Ewa Deelman², Yves Robert³, and
Uwe Schwiegelshohn⁴

1 University of Hawaii at Manoa, US, henric@hawaii.edu

2 University of Southern California, Marina del Rey, US, deelman@isi.edu

3 ENS Lyon, FR, & University of Tennessee – Knoxville, US,
Yves.Robert@ens-lyon.fr

4 TU Dortmund, DE, uwe.schwiegelshohn@udo.edu

Abstract

Large-scale systems face two main challenges: failure management and energy management. Failure management, the goal of which is to achieve resilience, is necessary because a large number of hardware resources implies a large number of failures during the execution of an application. Energy management, the goal of which is to optimize power consumption and to handle thermal issues, is also necessary due to both monetary and environmental constraints since typical applications executed in HPC and/or cloud environments will lead to large power consumption and heat dissipation due to intensive computation and communication workloads.

The main objective of this Dagstuhl seminar was to gather two communities: (i) system-oriented researchers who study high-level resource-provisioning policies, pragmatic resource allocation and scheduling heuristics, novel approaches for designing and deploying systems software infrastructures, and tools for monitoring/measuring the state of the system; and (ii) algorithm-oriented researchers, who investigate formal models and algorithmic solutions for resilience and energy efficiency problems. Both communities focused around workflow applications during the seminar, and discussed various issues related to the efficient, resilient, and energy efficient execution of workflows in distributed platforms.

This report provides a brief executive summary of the seminar and lists all the presented material.

Seminar July 6–10, 2015 – <http://www.dagstuhl.de/15281>

1998 ACM Subject Classification E.1 Data Structures, C.2.4 Distributed Systems, C.1.4 Parallel Architectures

Keywords and phrases Fault tolerance, Resilience, Energy efficiency, Distributed and high performance computing, Scheduling, Workflows

Digital Object Identifier 10.4230/DagRep.5.7.1



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

Algorithms and Scheduling Techniques to Manage Resilience and Power Consumption in Distributed Systems, *Dagstuhl Reports*, Vol. 5, Issue 7, pp. 1–21

Editors: Henri Casanova, Ewa Deelman, Yves Robert, and Uwe Schwiegelshohn



DAGSTUHL
REPORTS

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


1 Executive Summary

Henri Casanova

Ewa Deelman

Yves Robert

Uwe Schwiegelshohn

License  Creative Commons BY 3.0 Unported license

© Henri Casanova, Ewa Deelman, Yves Robert, and Uwe Schwiegelshohn

Many computer applications are executed on large-scale systems that comprise many hardware components, such as clusters that can be federated into distributed cloud computing or grid computing platforms. The owners/managers of these systems face two main challenges: failure management and energy management.

Failure management, the goal of which is to achieve resilience, is necessary because a large number of hardware resources implies a large number of failures during the execution of an application. While hardware resources can be made more reliable via the use of redundancy, this redundancy increases component cost. As a result, systems deployed within budget constraints must be built from unreliable components, that have a finite Mean Time Between Failure (MTBF), i.e., commercial-of-the-shelf components. For instance, a failure would occur every 50 minutes in a system with one million components, even if the MTBF of a single component is as large as 100 years.

Energy management, the goal of which is to optimize power consumption and to handle thermal issues, is also necessary due to both monetary and environmental constraints. While in today's systems, processors are the most power-consuming components, it is anticipated that in future distributed systems, the power dissipated to perform communications and I/O transfers will make up a much larger share of the overall energy consumption. In fact, the relative cost of communication is expected to increase dramatically, both in terms of latency/overhead and of consumed energy. Consequently, the computation and communication workloads of typical applications executed in HPC and/or cloud environments will lead to large power consumption and heat dissipation.

These two challenges, resilience and energy efficiency, are currently being studied by many researchers. Some of these researchers come from a “systems” culture, and investigate in particular systems design and management strategies that enhance resilience and energy efficiency. These strategies include high-level resource-provisioning policies, pragmatic resource allocation and scheduling heuristics, novel approaches for designing and deploying systems software infrastructures, and tools for monitoring/measuring the state of the system. Other researchers come from an “algorithms” culture. They investigate formal definitions of resilience and energy efficiency problems, relying on system models of various degrees of accuracy and sophistication, and aiming to obtain strong complexity results and algorithmic solutions for solving these problems. These two communities are quite often separated in the scientific literature and in the field. Some of the pragmatic solutions developed in the former community appear algorithmically weak to the latter community, while some of the algorithmic solutions developed by the latter community appear impractical to the former community. Furthermore, the separation of application and system platform due to ubiquitous resource virtualization layers also interferes with an effective cooperation of algorithmic and system management methods, and in particular to handle resiliency and energy efficiency. To move forward, more interaction and collaboration is needed between the systems and the algorithms communities, an observation that was made very clear during

the discussions in the predecessor Dagstuhl seminar ¹.

The broader challenge faced by systems and algorithms designer is that the optimization metrics of interest (resilience, power consumption, heat distribution, performance) are intimately related. For instance, high volatility in power consumption due to the use of dynamic frequency and voltage scaling (DFVS) is known to lead to thermal hotspots in a datacenter. Therefore, the datacenter must increase the safety margin for their cooling system to handle these hotspots. As a result, the power consumed by the cooling system is increased, possibly increasing the overall power consumption of the whole system, even though the motivation for using DVFS in the first place was to reduce power consumption! When resilience is thrown into the mix, then the trade-offs between the conflicting resilience, performance, and energy goals become even more intertwined. Adding fault-tolerance to a system, for instance, by using redundant computation or by periodically saving the state of the system to secondary storage, can decrease performance and almost always increases hardware resource requirements and thus power consumption. The field is rife with such conundrums, which must be addressed via systems and algorithms techniques used in conjunction. In this seminar, we have brought together researchers and practitioners from both the systems and the algorithms community, so as to foster fruitful discussions of these conundrums, many of which were touched upon in the predecessor seminar but by no means resolved.

To provide a clear context, the seminar focused around workflow applications. Workflows correspond to a broad and popular model of computation in which diverse computation tasks (which many themselves follow arbitrary models of computation) are interconnected via control and data dependencies. They have become very popular in many domains, ranging from scientific to datacenter applications, and share similar sets of challenges and current solutions. Part of the motivation of using workflows, and thus to develop workflow management systems and algorithms, is that they make it possible to describe complex and large computations succinctly and portably. Most of the invited seminar participants have worked and are currently working on issues related to the efficient, resilient, and energy efficient execution of workflows in distributed platforms. They thus provide an ideal focal and unifying theme for the seminar.

A number of workflow tools is available to aid the users in defining and executing workflow applications. While these tools are thus designed primarily to support the end user, they are in fact ideal proving grounds for implementing novel systems and algorithms techniques to aim at optimizing performance, resilience, and energy efficiency. Therefore, these tools provide a great opportunity to enhance both the application and the software infrastructure to meet both the needs of the end users and of the systems owners/managers. These goals are very diverse and, as we have seen above, intertwined, so that re-designing algorithms and systems to meet these goals is a difficult proposition (again, higher resilience often calls for redundant computations and/or redundant communication, which in turn consumes extra power and can reduce performance). In a broad sense, we are facing complex multi-criteria optimization problems that must be (i) formalized in a way that is cognizant of the practical systems constraints and hardware considerations; (ii) solved by novel algorithms that are both fast (so that they can be used in an on-line manner) and robust (so that they can tolerated wide ranges of scenarios with possibly inaccurate information).

The goal of this seminar was to foster discussions on, and articulate novel and promising directions for addressing the challenges highlighted above. International experts in the field

¹ Dagstuhl Seminar 13381, *Algorithms and Scheduling Techniques for Exascale Systems* (2013), organized by Henri Casanova, Yves Robert and Uwe Schwiegelshohn (<http://www.dagstuhl.de/13381>).

have investigated how to approach (and hopefully at least partially address) the challenges that algorithms and system designers face due to frequent failures and energy usage constraints. More specifically, the seminar has addressed the following topics:

- Multi-criteria optimization problems as applicable to fault-tolerance / energy management
- Resilience techniques for HPC and cloud systems
- Robust and energy-aware distributed algorithms for resource scheduling and allocation in large distributed systems.
- Application-specific approaches for fault-tolerance and energy management, with a focus on workflow-based applications

Although the presentations at the seminar were very diverse in scope, ranging from practice to theory, an interesting observation is that many works do establish strong links between practice (e.g., particular applications, programming models) and theory (e.g., abstract scheduling problems and results). In particular, it was found that workflow applications, far from being well-understood, in fact give rise to a range of interrelated and interesting practical and theoretical problems that must be solved conjointly to achieve efficiency at large scale. Estimating task weights, scheduling with uncertainties, mapping at scale, remapping after failures, trading performance and energy, these are a few challenges that have been discussed at length during the seminar. Such observations make it plain that forums that blends practice and theory, as is the case with this seminar, are very much needed.

The seminar brought together 41 researchers from Austria, France, Germany, Japan, Netherlands, New Zealand, Poland, Portugal, Spain, Sweden, Switzerland, UK and USA, with interests and expertise in different aspect of parallel and distributed computing. Among participants there was a good mix of senior researchers, junior researchers, postdoctoral researchers, and Ph.D. students. Altogether there were 29 presentations over the 5 days of the seminar, organized in morning and late-afternoon sessions. The program was as usual a compromise between allowing sufficient time for participants to present their work, while also providing unstructured periods that were used by participants to pursue ongoing collaborations as well as to foster new ones. The feedback provided by the participants show that the goals of the seminar, namely to circulate new ideas and create new collaborations, were met to a large extent.

The organizers and participants wish to thank the staff and the management of Schloss Dagstuhl for their assistance and support in the arrangement of a very successful and productive event.

2 Table of Contents

Executive Summary

<i>Henri Casanova, Ewa Deelman, Yves Robert, and Uwe Schwiegelshohn</i>	2
---	---

Overview of Talks

Heterogeneous supercomputing systems power-aware scheduling and resiliency <i>Sadaf Alam</i>	7
Analyzing real cluster data for formulating robust allocation algorithms in cloud platforms <i>Olivier Beaumont</i>	7
Which verification for soft error detection? <i>Anne Benoit</i>	8
Practical foundations for resilient applications <i>George Bosilca</i>	8
Task resource consumption prediction for scientific applications and workflows <i>Rafael Ferreira da Silva</i>	8
Cost constrained, static and dynamic workflow scheduling on IaaS clouds <i>Michael Gerhards</i>	9
Resilient multigrid solvers <i>Dominik Goeddeke</i>	10
Bridging the gap between performance and bounds of cholesky factorization on heterogeneous platforms <i>Julien Herrmann</i>	10
Accurately measuring MPI collectives with synchronized clocks <i>Sascha Hunold</i>	10
Energy-aware scheduling for task-based applications in distributed platforms <i>Fredy Juarez</i>	11
Platforms to fit and platforms to experiment <i>Katarzyna Keahey</i>	11
Some thoughts about extreme scaling and power <i>Dieter Kranzlmüller</i>	13
Communication lower bounds for matrix-matrix multiplication <i>Julien Langou</i>	13
Workflow scheduling and optimization on clouds <i>Maciej Malawski</i>	13
Scheduling trees of malleable tasks for sparse linear algebra <i>Loris Marchal</i>	14
The inevitable end of Moore's law beyond Exascale will result in data and HPC convergence <i>Satoshi Matsuoka</i>	14
Lazy shadowing: a power-aware resilience scheme for extreme scale systems <i>Rami Melhem</i>	15

Modelling Matlab/Simulink periodic applications using Synchronous DataFlow Graphs	
<i>Alix Munier</i>	15
HAEC – Highly Adaptive Energy-Efficient Computing	
<i>Wolfgang E. Nagel</i>	15
Application-aware approaches to resilience at extreme scales	
<i>Manish Parashar</i>	16
Cost-efficient resource management for scientific workflows on the cloud	
<i>Ilia Pietri</i>	16
A decade of modelling parallel computing systems	
<i>Sabri Pllana</i>	17
Dynamic node replacement and adaptive scheduling for fault tolerance	
<i>Suraj Prabhakaran</i>	17
Hybrid clouds: beyond classical buy-or-lease decisions	
<i>Stephan Schlagkamp</i>	18
Duplicate-free state-space model for optimal task scheduling	
<i>Oliver Sinnen</i>	18
Energy-aware task management for heterogeneous low-power systems	
<i>Leonel Sousa</i>	19
From static scheduling towards understanding uncertainty	
<i>Andrei Tchernykh</i>	19
Quantifying resiliency in the extreme scale hpc co-design space	
<i>Jeffrey S. Vetter</i>	19
Voltage overscaling algorithms for energy-efficient computations with timing errors	
<i>Frédéric Vivien</i>	20
Participants	21

3 Overview of Talks

3.1 Heterogeneous supercomputing systems power-aware scheduling and resiliency


Sadaf Alam (CSCS – Swiss National Supercomputing Centre, CH)

License  Creative Commons BY 3.0 Unported license
© Sadaf Alam

Swiss National Supercomputing Centre (CSCS) operates a highly energy efficient, Petscale hybrid Cray XC30 system, which has been in operation since the beginning of 2014. The system has a wide variety of user level and system logging power management and measurement tools that allow for fine grain power measurements of applications. However, the current scheduling and resource management scheme does not take into consideration the power requirements of applications. There is a potential for optimizing operating cost by carefully scheduling jobs based on an additional resource management parameter in addition to the node requirements and wall clock time. An analysis of system and job submission logs could provide insight into power aware resource management of the system. Likewise, resiliency on heterogeneous nodes have additional dimensions due to different types of execution units, memory as well as I/O. Correlating different error codes with failures allow us to detect, isolate and resolve issues with minimum impact to operations.

3.2 Analyzing real cluster data for formulating robust allocation algorithms in cloud platforms

Olivier Beaumont (INRIA – Bordeaux, FR)

License  Creative Commons BY 3.0 Unported license
© Olivier Beaumont
Joint work of Beaumont, Olivier; Eyraud-Dubois, Lionel; Lorenzo-del-Castillo, Juan-Angel

A problem commonly faced in Computer Science research is the lack of real usage data that can be used for the validation of algorithms. This situation is particularly true and crucial in Cloud Computing. The privacy of data managed by commercial Cloud infrastructures, together with their massive scale, makes them very uncommon to be available to the research community. Due to their scale, when designing resource allocation algorithms for Cloud infrastructures, many assumptions must be made in order to make the problem tractable.

This talk provides an analysis of a cluster data trace recently released by Google and focuses on a number of questions which have not been addressed in previous studies. In particular, we describe the characteristics of job resource usage in terms of dynamics (how it varies with time), of correlation between jobs (identify daily and/or weekly patterns), and correlation inside jobs between the different resources (dependence of memory usage on CPU usage) and at failures (are they independent or correlated). From this analysis, we propose a way to formalize the allocation problem on such platforms, which encompasses most job features from the trace with a small set of parameters.

3.3 Which verification for soft error detection?

Anne Benoit (ENS Lyon, FR)


License  Creative Commons BY 3.0 Unported license
© Anne Benoit

Joint work of Bautista-Gomez Leonardo; Benoit, Anne; Cavelan, Aurélien; Raina, Saurabh K.; Robert, Yves; Sun, Hongyang

Many methods are available to detect silent errors in high-performance computing (HPC) applications. Each comes with a given cost and recall (fraction of all errors that are actually detected). The main contribution of this paper is to show which detector(s) to use, and to characterize the optimal computational pattern for the application: how many detectors of each type to use, together with the length of the work segment that precedes each of them. We conduct a comprehensive complexity analysis of this optimization problem, showing NP-completeness and designing an FPTAS (Fully Polynomial-Time Approximation Scheme). On the practical side, we provide a greedy algorithm whose performance is shown to be close to the optimal for a realistic set of evaluation scenarios.

3.4 Practical foundations for resilient applications


George Bosilca (University of Tennessee, US)

License  Creative Commons BY 3.0 Unported license
© George Bosilca

Despite the fact that faults are accepted as normal occurrences, the existing programming paradigms are lacking methodologies and tools to deal with faults in a holistic manner. This talk covers an extension to the MPI paradigm allowing applications or libraries to provide transparent user level fault management solutions. Based on such solutions we describe several mixed solution, encompassing both traditional checkpoint/restart and application-specific approaches, to minimize the resilience overhead. Additionally, we introduce theoretical models to evaluate such mixed resilience models executed on particular hardware environments in order to assess the benefits and overhead of resilient applications at exascale.

3.5 Task resource consumption prediction for scientific applications and workflows

Rafael Ferreira da Silva (USC – Marina del Rey, US)

License  Creative Commons BY 3.0 Unported license
© Rafael Ferreira da Silva

Estimates of task runtime, disk space usage, and memory consumption, are commonly used by scheduling and resource provisioning algorithms to support efficient and reliable scientific application executions. Such algorithms often assume that accurate estimates are available, but such estimates are difficult to generate in practice. In this work, we first profile real scientific applications and workflows, collecting fine-grained information such as process I/O, runtime, memory usage, and CPU utilization. We then propose a method to automatically characterize task requirements based on these profiles. Our method estimates task runtime, disk space, and peak memory consumption. It looks for correlations between the parameters

of a dataset, and if no correlation is found, the dataset is divided into smaller subsets using the statistical recursive partitioning method and conditional inference trees to identify patterns that characterize particular behaviors of the workload. We then propose an estimation process to predict task characteristics of scientific applications based on the collected data. For scientific workflows, we propose an online estimation process based on the MAPE-K loop, where task executions are monitored and estimates are updated as more information becomes available. Experimental results show that our online estimation process results in much more accurate predictions than an offline approach, where all task requirements are estimated prior to workflow execution.

3.6 Cost constrained, static and dynamic workflow scheduling on IaaS clouds

Michael Gerhards (FH Aachen – Jülich, DE)

License © Creative Commons BY 3.0 Unported license
© Michael Gerhards

This talk presents a novel cloud-aware workflow scheduling approach. Scientific computing infrastructures of the last years were dominated by grids forming resource pools provided by independent organizations. These resources were provided and accessed following a set of rules that finally are the foundation of the virtual organization (VO). The cloud computing paradigm offers alternative computing infrastructures such as Amazon EC2 and Microsoft Azure. While both paradigms focus on characteristics such as broad network access, resource pooling, and measured service, cloud computing additionally prioritizes on-demand self-service and rapid elasticity mostly offered by a pay-as-you-go business model. With respect to scheduling, these cloud specific characteristics introduces new opportunities and challenges.

Large-scale scientific workflow applications composed of inter-dependent computational tasks are an important group of applications. Since the scheduling of these workflows is NP-complete, practical solutions will have to sacrifice optimality for the sake of efficiency. The literature distinguishes between two types of scheduling mechanisms, namely static scheduling and dynamic scheduling. In the static scheduling mechanism, all decisions are made offline before starting the execution of an application. Here, the question arises, how the knowledge needed to perform static scheduling is acquired. However, the workflow's runtime behaviors might differ from offline-planning. Hence, the static resource allocation plan might become uneconomic or might even fail in meeting a deadline. In the contrary, in the dynamic scheduling mechanism, all decisions are made at runtime under consideration of the system feedback. Hence, dynamic scheduling provides a more adaptive solution. However, the commonly implemented Master-Slave model assumes the existence of slave resources and does therefore neglect the problem of selecting a cost optimized set of resources in the cloud.

The outlined problems of both scheduling mechanisms motivate the main research hypothesis of this thesis: "How to combine the benefits of cloud-aware static and dynamic workflow scheduling mechanisms consistently in one solution?" More precisely, the following central research objectives are addressed:

- Identification of cost reduction strategies for workflow scheduling on pay-as-you-go clouds
- Estimation of task runtime profiles for heterogeneous resources based on provenance data
- Deriving of a flow pattern based polynomial complex static workflow scheduling algorithm
- Deriving of a resource allocation plan based low complex dynamic workflow scheduling algorithm

3.7 Resilient multigrid solvers

Dominik Goeddeke (TU Dortmund, DE)


License  Creative Commons BY 3.0 Unported license
© Dominik Goeddeke

Joint work of Altenbernd, Mirco; Goeddeke, Dominik; Ribbrock, Dick

We discuss fault tolerance schemes for data loss in multigrid solvers, that essentially combine ideas of checkpoint-restart with algorithm-based fault tolerance. We first demonstrate and prove that multigrid is self-stabilising with respect to single faults. With increasing fault rates however, checkpoints are required. To improve efficiency compared to conventional checkpointing, we exploit the inherent data compression of the multigrid hierarchy, and relax the synchronicity requirement through a local failure local recovery approach. We experimentally identify the root cause of convergence degradation in the presence of data loss using smoothness considerations. While these techniques primarily aim at the case of lost nodes, towards the end of the talk we also discuss a novel black-box approach based on the Full Approximation Scheme (FAS) that encapsulates all bitflips in the smoother and protects the traversal through the hierarchy with checksums.

3.8 Bridging the gap between performance and bounds of cholesky factorization on heterogeneous platforms

Julien Herrmann (ENS Lyon, FR)

License  Creative Commons BY 3.0 Unported license
© Julien Herrmann

Joint work of Agullo, Emmanuel; Beaumont, Olivier; Eyraud-Dubois, Lionel; Herrmann, Julien; Kumar, Suraj; Marchal, Loris; Thibault, Samuel

We consider the problem of allocating and scheduling dense linear application on fully heterogeneous platforms made of CPUs and GPUs. More specifically, we focus on the Cholesky factorization since it exhibits the main features of such problems. Indeed, the relative performance of CPU and GPU highly depends on the sub-routine: GPUs are for instance much more efficient to process regular kernels such as matrix-matrix multiplications rather than more irregular kernels such as matrix factorization. In this context, one solution consists in relying on dynamic scheduling and resource allocation mechanisms such as the ones provided by PaRSEC or StarPU. In this paper we analyze the performance of dynamic schedulers based on both actual executions and simulations, and we investigate how adding static rules based on an offline analysis of the problem to their decision process can indeed improve their performance, up to reaching some improved theoretical performance bounds which we introduce.

3.9 Accurately measuring MPI collectives with synchronized clocks

Sascha Hunold (TU Wien, AT)

License  Creative Commons BY 3.0 Unported license
© Sascha Hunold

We consider the problem of accurately measuring the time to complete an MPI collective operation, as the result strongly depends on how the time is measured. Our goal is to develop an experimental method that allows for reproducible measurements of MPI collectives. When

executing large parallel codes, MPI processes are often skewed in time when entering a collective operation. However, for the sake of reproducibility, it is a common approach to synchronize all processes before they call the MPI collective operation. We therefore take a closer look at two commonly used process synchronization schemes: (1) relying on *MPIBarrier* or (2) applying a window-based scheme using a common global time. We analyze both schemes experimentally and show the pros and cons of each approach. As window-based schemes require the notion of global time, we thoroughly evaluate different clock synchronization algorithms in various experiments. We also propose a novel clock synchronization algorithm that combines two advantages of known algorithms, which are (1) taking the inherent clock drift into account and (2) using a tree-based synchronization scheme to reduce the synchronization duration.

3.10 Energy-aware scheduling for task-based applications in distributed platforms

Fredy Juarez (Barcelona Supercomputing Center, ES)

License © Creative Commons BY 3.0 Unported license
© Fredy Juarez

Joint work of Badia, Rosa; Ejarque, Jorge; Juarez, Fredy

Green Computing is a recent trend in computer science which tries to reduce the energy consumption and carbon footprint produced by computers. One of the methods to reduce this consumption is providing scheduling policies for taking into account not only the processing time but also the energy consumption. We propose a real-time dynamic scheduling system to efficiently execute task-based applications on Distributed Computing platforms such as Cluster, Grids and Clouds. The proposed scheduler minimizes a multi-objective function which combines the energy-consumption and execution time according to the energy-performance importance factor provided by the user or cloud provider. Due to the time limitation required for run-time schedulers, the implementation of large-time optimization algorithms are not suitable. Therefore, we combine a set of heuristic rules plus the resource allocator algorithm to get good real-time scheduling solutions in an affordable time scale.

3.11 Platforms to fit and platforms to experiment

Katarzyna Keahey (Argonne National Laboratory, US)

License © Creative Commons BY 3.0 Unported license
© Katarzyna Keahey

The data processing needs of applications are undergoing a significant change, increasingly emphasizing on-demand availability and real-time or simply predictable response time, driven by availability of dynamic data streams from experimental instruments, social networks, and specialized sensing devices; a new disruptive factor promising to be the primary driver of discovery in environmental and other sciences over the next decade. Applications of this type radically change what is required of scientific resources. Where before a resource would provide a static offering and require an application to adapt itself to hardware and configuration trade-offs offered by the resource, now the requirement for on-demand availability is increasingly requiring a resource to be able to adapt itself the requirements of the application in time

to service time-sensitive requests. In other words, resource providers are asked to offer a programmable platform, that can adapt itself to environment and hardware requirements of an application.

This shifts emphasis of many questions we used to ask about software. Where before adapting the application was the main focus of optimization, now time is spent in developing technologies that adapt the platform. Where before available platforms could be explored manually and at a coarse grain, now automated discovery and fine-grained understanding of the platform's capabilities become of paramount importance. And finally, once a platform becomes dynamic its ability to adapt may be used during a run's lifecycle as well as before providing dynamic support for malleable applications. A factor essential to achieving all this is application's ability to comprehensively define its needs.

This talk describes two types of technologies essentially to create a programmable platform: "platform to fit" technologies, that adapt resources to the needs of the application, and "fitting datacenter" technologies that ensure that the resource is presented such that it forms good building blocks for such adaptability. The critical element of a "fitting datacenter" is a resource manager that provides resource leases to users and applications that offer a significant degree of programmability both in terms of environment configuration and the shape of a resource slot. A key element of such resource manager are containers, implemented via a variety of technologies such as virtual machines (e.g., KVM or Xen) or technologies based on OS-based constructs such as Docker or LXE. To accommodate both on-demand availability and high utilization such resource manager has to find ways to efficiently interleave on-demand leases with the demands of batch and high throughput computing codes. The "platform to fit" leverages on-demand availability of a variety of resources to provide adaptability in the platform. Resource adaptability can happen in the compute dimension, where integrating additional resources by deploying on-demand VMs can keep response times stable when dealing with peak traffic from dynamic data streams and can be regulated by a variety of policies. Similarly available storage capacity can be adapted to the needs of an application or a workload, but dynamically adding and releasing storage volumes given efficient filesystem supporting such operation. And finally the available I/O bandwidth can also be adapted by placing a caching system in front of storage offerings with different I/O bandwidths at different price points; this means that data could be managed transparently between fast disks during computation or periods of intensive access, and archival storage when periodically not used.

Experimental testbeds for computer science are critical for experimenting with these types of technologies. The new NSF-funded Chameleon testbed (www.chameleoncloud.org) has been built specifically to provide such experimental capabilities. It's hardware makup, consisting of 15,000 cores distributed over two sites connected by a 100Gbps network, and concentrated primarily in one homogenous partition, as well as 5 PB of storage provide a good canvas for Big Data and Big Compute experiments. Chameleon offers bare hardware reconfiguration allowing users to recreate as nearly as possible conditions available in their own lab and offers reproducibility. The testbed is currently in the Early User stage and will be publicly available by fall of 2015.

3.12 Some thoughts about extreme scaling and power

Dieter Kranzlmüller (LMU München, DE)

License  Creative Commons BY 3.0 Unported license
© Dieter Kranzlmüller

This talk discusses the (scheduling) problems on the supercomputer SuperMUC Phase 1 and 2, operating at the Leibniz Supercomputing Centre (LRZ). The issues include scalability, power, scheduling, and infrastructure issues.

3.13 Communication lower bounds for matrix-matrix multiplication


Julien Langou (University of Colorado Denver, US)

License  Creative Commons BY 3.0 Unported license
© Julien Langou

We consider communication lower bounds for matrix-matrix multiplication in the sequential case. Our new proof technique improves the known lower bound for the number of reads and writes in the sequential memory case.

3.14 Workflow scheduling and optimization on clouds

Maciej Malawski (AGH University of Science & Technology, Krakow, PL)

License  Creative Commons BY 3.0 Unported license
© Maciej Malawski

The area of my research includes the problem of workflow scheduling on IaaS clouds.

The first part of the talk focuses on data-intensive workflows, and addresses the problem of scheduling workflow ensembles under cost and deadline constraints in clouds. We developed a simulation model for handling file transfers between tasks, featuring the ability to dynamically calculate bandwidth and supporting a configurable number of replicas, thus allowing us to simulate various levels of congestion. The resulting model is capable of representing a wide range of storage systems available on clouds: from in-memory caches (such as memcached), to distributed file systems (such as NFS servers) and cloud storage (such as Amazon S3). Next, we propose and evaluate a novel scheduling algorithm that minimizes the number of transfers by taking advantage of data caching and file locality.

The second part of the talk covers other aspects of cloud workflow scheduling. They include problems with inaccurate run time estimates and task granularity. When addressing the problem of scheduling on multiple clouds, we developed optimization methods based on mixed integer programming. We have also performed experiments with performance evaluation of applications on clouds and developed a performance model of workflow for ISMOP flood prevention project.

3.15 Scheduling trees of malleable tasks for sparse linear algebra

Loris Marchal (ENS Lyon, FR)

License © Creative Commons BY 3.0 Unported license
© Loris Marchal

Joint work of Guermouche, Abdou; Marchal, Loris; Simon, Bertrand; Vivien, Frédéric

Scientific workloads are often described by directed acyclic task graphs. This is in particular the case for multifrontal factorization of sparse matrices – the focus of this talk – whose task graph is structured as a tree of parallel tasks. Prasanna and Musicus advocated using the concept of malleable tasks to model parallel tasks involved in matrix computations. In this powerful model each task is processed on a time-varying number of processors. Following Prasanna and Musicus, we consider malleable tasks whose speedup is p^α , where p is the fractional share of processors on which a task executes, and α ($0 < \alpha \leq 1$) is a task-independent parameter. Firstly, we use actual experiments on multicore platforms to motivate the relevance of this model for our application. Then, we study the optimal time-minimizing allocation proposed by Prasanna and Musicus using optimal control theory. We greatly simplify their proofs by resorting only to pure scheduling arguments. Building on the insight gained thanks to these new proofs, we extend the study to distributed (homogeneous or heterogeneous) multicore platforms. We prove the NP-completeness of the corresponding scheduling problem, and we then propose some approximation algorithms.

3.16 The inevitable end of Moore’s law beyond Exascale will result in data and HPC convergence

Satoshi Matsuoka (Tokyo Institute of Technology, JP)

License © Creative Commons BY 3.0 Unported license
© Satoshi Matsuoka

The so-called “Moore’s Law”, by which the performance of the processors will increase exponentially by factor of 4 every 3 years or so, is slated to be ending in 10–15 year timeframe due to the lithography of VLSIs reaching its limits around that time, and combined with other physical factors. This is largely due to the transistor power becoming largely constant, and as a result, means to sustain continuous performance increase must be sought otherwise than increasing the clock rate or the number of floating point units in the chips, i.e., increase in the FLOPS. The promising new parameter in place of the transistor count is the perceived increase in the capacity and bandwidth of storage, driven by device, architectural, as well as packaging innovations: DRAM-alternative Non-Volatile Memory (NVM) devices, 3-D memory and logic stacking evolving from VIAs to direct silicone stacking, as well as next-generation terabit optics and networks. The overall effect of this is that, the trend to increase the computational intensity as advocated today will no longer result in performance increase, but rather, exploiting the memory and bandwidth capacities will instead be the right methodology. However, such shift in compute-vs-data tradeoffs would not exactly be return to the old vector days, since other physical factors such as latency will not change. As such, performance modeling to account for the evolution of such fundamental architectural change in the post-Moore era would become important, as it could lead to disruptive alterations on how the computing system, both hardware and software, would be evolving towards the future.

3.17 Lazy shadowing: a power-aware resilience scheme for extreme scale systems

Rami Melhem (University of Pittsburgh, US)

License © Creative Commons BY 3.0 Unported license
© Rami Melhem

As the demand for high performance computing (HPC) and cloud computing accelerates, the underlying infrastructure is expected to ensure reliability while maintaining high performance and cost-effectiveness, even with multifold increase in the number of computing, storage and communication components. Current resilience approaches rely upon either time redundancy (re-execution after failure) or hardware redundancy (executing multiple copies). The first approach is subject to a significant delay while the second approach requires additional hardware and increases the energy consumption for a given service. In general, the trade-off between performance, fault-tolerance and power consumption calls for new frameworks which is energy and performance aware when dealing with failures. In this talk, Shadow Computing is introduced to address the above trade-off challenge by associating with each main process a shadow which executes at a reduced rate through either voltage/frequency scaling or by co-locating multiple shadows on the same processor. Adjusting the rate of execution enables a parameterized trade-off between response time, energy consumption and hardware redundancy.

3.18 Modelling Matlab/Simulink periodic applications using Synchronous DataFlow Graphs

Alix Munier (UPMC – Paris, FR)

License © Creative Commons BY 3.0 Unported license
© Alix Munier

Matlab/Simulink is a simulation tool widely used in an industrial context like automotive to design embedded electronic applications.

The aim of this talk is to prove that communications between Simulink periodic blocks can be described using a Synchronous Dataflow Graph (SDF in short). Therefore, the amounts of data exchanged between the blocks are modeled by closed formulas and are completely characterized without any simulation phase. The major consequence is that applications designed using Simulink can take advantage of the (semi)-automatic tools for optimized execution of a SDF on a many-core architecture.

3.19 HAEC – Highly Adaptive Energy-Efficient Computing

Wolfgang E. Nagel (TU Dresden, DE)


License © Creative Commons BY 3.0 Unported license
© Wolfgang E. Nagel

The visionary goal of the Collaborative Research Center HAEC (highly adaptive energy-efficient computing) is to research technologies to enable computing systems with high energy-efficiency without compromising on high performance. HAEC will concentrate on

researching larger server systems, from applications to hardware. To achieve the goal of an integrated approach of highly adaptive energy-efficient computing (HAEC), the problem is approached at all levels of technology involved, the hardware, the computer architecture, and operating system, the software modeling as well as the application modeling and runtime control levels. A novel concept, namely the HAEC Box, of how computers can be built by utilizing innovative ideas of optical and wireless chip-to-chip communication is explored. This would allow a new level of run-time adaptivity of future computers, creating a platform for flexibly adapting to the needs of the computing problem. The talk will describe the general approach and the implications on scheduling challenges in future systems.

3.20 Application-aware approaches to resilience at extreme scales

Manish Parashar (Rutgers University, US)

License  Creative Commons BY 3.0 Unported license
© Manish Parashar

Application resilience is a key challenge as we target exascale, and process/node failures represent an important class of failures that must be addressed. However, typical approaches for handling these failures, such as those based on terminating jobs and restarting them from the last stored checkpoints are not expected to scale to exascale. In this talk I will explore application driven approaches that use knowledge of the application and/or the methods used to reduce overheads due to fault tolerance for MPI-based parallel applications, and increase their resilience. Specifically, I will present approaches for online global and local recovery, implicitly coordinated checkpointing, and failure masking. I will also describe Fenix, a scalable framework for enabling online recovery from process/node/blade/cabinet failures. Finally, I will present evaluations of the developed approaches and of Fenix using the S3D combustion simulation running on the Titan Cray-XK7 production system at ORNL.

3.21 Cost-efficient resource management for scientific workflows on the cloud

Ilia Pietri (University of Manchester, UK)

License  Creative Commons BY 3.0 Unported license
© Ilia Pietri

Scientific workflows, which describe complex computational problems in many scientific fields, may be executed on high performance systems such as Clouds. Cloud providers now offer to users CPU provisioning as different combinations of CPU frequencies and prices. With the increasing number of options, the problem of choosing cost-efficient configurations is becoming more challenging. This talk discusses approaches to minimize the energy cost from the provider's perspective and the user monetary cost while achieving good performance in terms of execution time. It presents a performance model to estimate the workflow makespan under different configurations and frequency selection approaches to determine the CPU frequencies to operate the available resources and execute the workflow tasks, based on both the application and system characteristics.

3.22 A decade of modelling parallel computing systems

Sabri Pllana (Linnaeus University, Växjö, SE)

License © Creative Commons BY 3.0 Unported license
© Sabri Pllana

Joint work of Suraj Prabhakaran, Marcel Neumann, Felix Wolf

In mature engineering disciplines such as civil engineering, before an artefact (for instance a bridge) is built, the corresponding model is developed. We argue that a practice “model first then build the code” would benefit also software engineers. Since programming parallel systems is considered significantly more complex than programming sequential systems, the use of models in the context of parallel computing systems is of particular importance. In this talk we will highlight some of our models of parallel computing systems that we have developed in the last decade. We will first address various modelling aspects in the context of clusters of SMPs, continue thereafter with the Grid, and conclude this talk with heterogeneous computing systems.

3.23 Dynamic node replacement and adaptive scheduling for fault tolerance

Suraj Prabhakaran (TU Darmstadt, DE)


License © Creative Commons BY 3.0 Unported license
© Suraj Prabhakaran

Joint work of Neumann, Marcel; Prabhakaran, Suraj; Wolf, Felix

Batch systems traditionally support only static resource management wherein a job’s resource set is unchanged throughout execution. Node failures force the batch systems to restart affected jobs on a fresh allocation (typically from a checkpoint) or replace failed nodes with statically allocated spare nodes. As future systems are expected to have high failure rates, this solution leads to increased job restart overhead, additional long waiting times before job restart and excessive resource wastage. In this talk, we present an extension of the TORQUE/Maui batch system with dynamic resource management facilities that enable instant replacement of failed nodes to affected jobs without requiring a job restart. The proposed batch system supports all job types for scheduling – rigid, moldable, evolving and malleable. We present an algorithm for the combined scheduling of all job types and show how the unique features of various jobs and the scheduling algorithm can expedite node replacements. The overall expected benefit of this approach is a highly resilient cluster environment that ensures timely completion of jobs and maintains high throughput even under frequent node failures.

3.24 Hybrid clouds: beyond classical buy-or-lease decisions

Stephan Schlagkamp (TU Dortmund, DE)

License  Creative Commons BY 3.0 Unported license
© Stephan Schlagkamp

Joint work of Schlagkamp, Stephan; Schwiegelshohn, Uwe; Tchernykh, Andrei

An increasing number of companies favor hybrid clouds since this approach does not require vital company data to leave the premises while it still allows efficient handling of demand peaks. Those companies must determine the size of the private cloud component by considering cost and demand forecasts. We formulate an optimization problem addressing cost optimality of a hybrid cloud at strategic level. Furthermore, we present a simple method to calculate a cost optimal size of the private cloud given a statistical demand prediction. Such demand prediction respects unavoidable deviations in prediction. Moreover, we analyze robustness of our method against errors in cost and demand estimations.

3.25 Duplicate-free state-space model for optimal task scheduling

Oliver Sinnen (University of Auckland, NZ)

License  Creative Commons BY 3.0 Unported license
© Oliver Sinnen

Joint work of Orr, Michael; Sinnen, Oliver

The problem of task scheduling with communication delays ($P|prec, c_{ij}|C_{max}$) is NP-hard, and therefore solutions are often found using a heuristic method. However, an optimal schedule can be very useful in applications such as time critical systems, or as a baseline for the evaluation of heuristics. Branch-and-bound algorithms such as A* have previously been shown to be a promising approach to the optimal solving of this problem, using a state-space model which we refer to as exhaustive list scheduling. However, this model suffers from the possibility of producing high numbers of duplicate states. In this paper we define a new state-space model in which we divide the problem into two distinct subproblems: first we decide the allocations of all tasks to processors, and then we order the tasks on their allocated processors in order to produce a complete schedule. This two-phase state-space model offers no potential for the production of duplicates. An empirical evaluation shows that the use of this new state-space model leads to a marked reduction in the number of states considered by an A* search in many cases, particularly for task graphs with a high communication-to-computation ratio. With additional refinement, and the development of specialised pruning techniques, the performance of this state-space model could be further improved.

3.26 Energy-aware task management for heterogeneous low-power systems

Leonel Sousa (Technical University of Lisboa, PT)

License © Creative Commons BY 3.0 Unported license
© Leonel Sousa

This talk presents a framework for energy-aware task management in heterogeneous embedded platforms, which integrates a set of novel application-aware management mechanisms for efficient resource utilization, frequency scaling and task migration. These mechanisms rely on a new management concept, which promotes performance fairness among running tasks to attain energy savings, while respecting the target performance of parallel applications. The proposed framework integrates several components for accurate run-time monitoring and application self-reporting. Experimental results show that energy savings of up to 39% were achieved in a state-of-the-art embedded platform for a set of real-world SPEC CPU2006 and PARSEC benchmarks.

3.27 From static scheduling towards understanding uncertainty

Andrei Tchernykh (CICESE Research Center, US)

License © Creative Commons BY 3.0 Unported license
© Andrei Tchernykh

Clouds differ from previous computing environments in the way that they introduce a continuous uncertainty into the computational process. The uncertainty brings additional challenges to both end-users and resource providers. It requires waiving habitual computing paradigms, adapting current computing models to this evolution, and designing novel resource management strategies to mitigate uncertainty and handle it in an effective way. In this talk, we address scheduling algorithms for different scenarios of HPC, Grid and Cloud Infrastructures. We provide some theoretical and experimental bounds for static, dynamic and adaptive approaches. We discuss the role of uncertainty in the resource/service provisioning, investment, operational cost, programming models. We discuss several major sources of uncertainty: dynamic elasticity, dynamic performance changing, virtualization, loosely coupling application to the infrastructure, among many others.

3.28 Quantifying resiliency in the extreme scale hpc co-design space

Jeffrey S. Vetter (Oak Ridge National Laboratory, US)


License © Creative Commons BY 3.0 Unported license
© Jeffrey S. Vetter
Joint work of Meredith, Jeremy; Vetter, Jeffrey S.

A key capability for the co-design of extreme-scale HPC systems is the accurate modeling of performance, power, and resiliency. We have developed the Aspen performance modeling language that allows fast exploration of the holistic design space. Aspen is a domain specific language for structured analytical modeling of applications and architectures. Aspen specifies a formal grammar to describe an abstract machine model and describe an application's behaviors, including available parallelism, operation counts, data structures, and control

flow. Aspen is designed to enable rapid exploration of new algorithm and architectures. Recently, we have used Aspen to model application resiliency based on data vulnerability. To facilitate this vulnerability classification, it is important to have accurate, quantitative techniques that can be applied uniformly and automatically across real-world applications. Traditional methods cannot effectively quantify vulnerability, because they lack a holistic view to examine system resilience, and come with prohibitive evaluation costs. To attack this problem, we introduce a data-driven, practical methodology to analyze these application vulnerabilities using a novel resilience metric: the data vulnerability factor (DVF). DVF integrates knowledge from both the application and target hardware into the calculation. To calculate DVF, we extend the Aspen performance modeling language to provide a structured, fast modeling solution.

3.29 Voltage overscaling algorithms for energy-efficient computations with timing errors

Frédéric Vivien (ENS Lyon, FR)

License  Creative Commons BY 3.0 Unported license

© Frédéric Vivien

Joint work of Cavelan, Aurélien; Robert, Yves; Sun, Hongyang; Vivien, Frédéric

In this work, we discuss several scheduling algorithms to execute tasks with voltage overscaling. Given a frequency to execute the tasks, operating at a voltage below threshold leads to significant energy savings but also induces timing errors. A verification mechanism must be enforced to detect these errors. Contrarily to fail-stop or silent errors, timing errors are deterministic (but unpredictable). For each task, the general strategy is to select a voltage for execution, to check the result, and to select a higher voltage for re-execution if a timing error has occurred, and so on until a correct result is obtained. Switching from one voltage to another incurs a given cost, so it might be efficient to try and execute several tasks at the current voltage before switching to another one. Determining the optimal solution turns out to be unexpectedly difficult. However, we provide the optimal algorithm for a single task and for a chain of tasks, the optimal algorithm when there are only two voltages, and the optimal level algorithm for several tasks, where a level algorithm is defined as an algorithm that executes all remaining tasks when switching to a given voltage. Furthermore, we show that the optimal level algorithm is in fact globally optimal (among all possible algorithms) when voltage switching costs are linear. Finally, we report a set of simulations to assess the potential gain of voltage overscaling algorithms.

Participants

- Sadaf Alam
CSCS – Lugano, CH
- Olivier Beaumont
INRIA – Bordeaux, FR
- Anne Benoit
ENS – Lyon, FR
- George Bosilca
University of Tennessee,
Knoxville, US
- Henri Casanova
University of Hawaii at Manoa –
Honolulu, US
- Ewa Deelman
USC – Marina del Rey, US
- Rafael Ferreira da Silva
USC – Marina del Rey, US
- Carsten Franke
ABB Corporate Research –
Baden-Dättwil, CH
- Michael Gerhards
FH Aachen – Jülich, DE
- Dominik Göddeke
Universität Stuttgart, DE
- Julien Herrmann
ENS – Lyon, FR
- Sascha Hunold
TU Wien, AT
- Fredy Juarez
Barcelona Supercomputing
Center, ES
- Kate Keahey
Argonne National Laboratory, US
- Thilo Kielmann
VU University Amsterdam, NL
- Dieter Kranzlmüller
LMU München, DE
- Julien Langou
Univ. of Colorado – Denver, US
- Maciej Malawski
AGH University of Science &
Technology – Krakow, PL
- Loris Marchal
ENS – Lyon, FR
- Satoshi Matsuoka
Tokyo Institute of Technology, JP
- Rami Melhem
University of Pittsburgh, US
- Bernd Mohr
Jülich Supercomputing
Centre, DE
- Alix Munier
UPMC – Paris, FR
- Jaroslaw Nabrzyski
Univ. of Notre Dame, US
- Wolfgang E. Nagel
TU Dresden, DE
- vManish Parashar
Rutgers Univ. – Piscataway, US
- Ilia Pietri
University of Manchester, GB
- Sabri Pllana
Linnaeus University – Växjö, SE
- Suraj Prabhakaran
TU Darmstadt, DE
- Padma Raghavan
Pennsylvania State University –
University Park, US
- Yves Robert
ENS – Lyon, FR
- Stephan Schlagkamp
TU Dortmund, DE
- Uwe Schwiegelshohn
TU Dortmund, DE
- Oliver Sinnen
University of Auckland, NZ
- Renata Slota
AGH University of Science &
Technology – Krakow, PL
- Veronika Sonigo
University of Franche-Comté –
Besançon, FR
- Leonel Sousa
Technical Univ. of Lisboa, PT
- Andrei Tchernykh
CICESE Research Center, US
- Jeffrey S. Vetter
Oak Ridge National Lab., US
- Frédéric Vivien
ENS – Lyon, FR
- Felix Wolf
TU Darmstadt, DE



The Constraint Satisfaction Problem: Complexity and Approximability

Edited by

Andrei A. Bulatov¹, Venkatesan Guruswami², Andrei Krokhin³,
and Dániel Marx⁴

1 Simon Fraser University – Burnaby, CA, abulatov@sfu.ca

2 Carnegie Mellon University, US, guruswami@cmu.edu

3 Durham University, GB, andrei.krokhin@durham.ac.uk

4 Hungarian Academy of Sciences – Budapest, HU, dm Marx@cs.bme.hu

Abstract

During the past two decades, an impressive array of diverse methods from several different mathematical fields, including algebra, logic, mathematical programming, probability theory, graph theory, and combinatorics, have been used to analyze both the computational complexity and approximability of algorithmic tasks related to the constraint satisfaction problem (CSP), as well as the applicability/limitations of algorithmic techniques. This research direction develops at an impressive speed, regularly producing very strong and general results. The Dagstuhl Seminar 15301 “The Constraint Satisfaction Problem: Complexity and Approximability” was aimed at bringing together researchers using all the different techniques in the study of the CSP, so that they can share their insights obtained during the past three years. This report documents the material presented during the course of the seminar.

Seminar July 19–24, 2015 – <http://www.dagstuhl.de/15301>

1998 ACM Subject Classification F.2.0 Analysis of Algorithms and Problem Complexity – General

Keywords and phrases Constraint satisfaction problem (CSP), Computational complexity, CSP dichotomy conjecture, Hardness of approximation, Unique games conjecture, Fixed-parameter tractability, Descriptive complexity, Universal algebra, Logic, Decomposition methods

Digital Object Identifier 10.4230/DagRep.5.7.22

Edited in cooperation with Alexandr Kazda


1 Executive Summary

Andrei A. Bulatov

Venkatesan Guruswami

Andrei Krokhin

Dániel Marx

License  Creative Commons BY 3.0 Unported license

© Andrei A. Bulatov, Venkatesan Guruswami, Andrei Krokhin, and Dániel Marx

The *constraint satisfaction problem*, or CSP in short, provides a unifying framework in which it is possible to express, in a natural way, a wide variety of computational problems dealing with mappings and assignments, including satisfiability, graph colorability, and systems of equations. The CSP framework originated 25–30 years ago independently in artificial intelligence, database theory, and graph theory, under three different guises, and



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

The Constraint Satisfaction Problem: Complexity and Approximability, *Dagstuhl Reports*, Vol. 5, Issue 7, pp. 22–41

Editors: Andrei A. Bulatov, Venkatesan Guruswami, Andrei Krokhin, and Dániel Marx



Dagstuhl Reports

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

it was realised only in the late 1990s that these are in fact different faces of the same fundamental problem. Nowadays, the CSP is extensively used in theoretical computer science, being a mathematical object with very rich structure that provides an excellent laboratory both for classification methods and for algorithmic techniques, while in AI and more applied areas of computer science this framework is widely regarded as a versatile and efficient way of modelling and solving a variety of real-world problems, such as planning and scheduling, software verification and natural language comprehension, to name just a few. An instance of CSP consists of a set of variables, a set of values for the variables, and a set of constraints that restrict the combinations of values that certain subsets of variables may take. Given such an instance, the possible questions include (a) deciding whether there is an assignment of values to the variables so that every constraint is satisfied, or optimising such assignments in various ways, (b) counting satisfying assignments, exactly or approximately, or (c) finding an assignment satisfying as many constraints as possible. There are many important modifications and extensions of this basic framework, e.g. those that deal with valued or global constraints.

Constraint satisfaction has always played a central role in computational complexity theory; appropriate versions of CSPs are classical complete problems for most standard complexity classes. CSPs constitute a very rich and yet sufficiently manageable class of problems to give a good perspective on general computational phenomena. For instance, they help to understand which mathematical properties make a computational problem tractable (in a wide sense, e.g. polynomial-time solvable or non-trivially approximable, fixed-parameter tractable or definable in a weak logic). It is only natural that CSPs play a role in many high-profile conjectures in complexity theory, exemplified by the Dichotomy Conjecture of Feder and Vardi and the Unique Games Conjecture of Khot.

The recent flurry of activity on the topic of the seminar is witnessed by three previous Dagstuhl seminars, titled “Complexity of constraints” (06401) and “The CSP: complexity and approximability” (09441, 12541), that were held in 2006, 2009, and 2012 respectively. This seminar was a follow-up to the 2009 and 2012 seminars. Indeed, the exchange of ideas at the 2009 and 2012 seminars has led to new ambitious research projects and to establishing regular communications channels, and there is a clear potential of a further systematic interaction that will keep on cross-fertilizing the areas and opening new research directions. The 2015 seminar brought together forty three researchers from different highly advanced areas of constraint satisfaction and involved many specialists who use universal-algebraic, combinatorial, geometric and probabilistic techniques to study CSP-related algorithmic problems. The participants presented, in 28 talks, their recent results on a number of important questions concerning the topic of the seminar. One particular feature of this seminar is a significant increase in the number of talks involving multiple subareas and approaches within its research direction – a definite sign of the growing synergy, which is one of the main goals of this series of seminars.

Concluding Remarks and Future Plans. The seminar was well received as witnessed by the high rate of accepted invitations and the great degree of involvement by the participants. Because of the multitude of impressive results reported during the seminar and the active discussions between researchers with different expertise areas, the organisers regard this seminar as a great success. With steadily increasing interactions between such researchers, we foresee a new seminar focussing on the interplay between different approaches to studying the complexity and approximability of the CSP. Finally, the organisers wish to express their gratitude to the Scientific Directors of the Dagstuhl Centre for their support of the seminar.

Description of the Topics of the Seminar

Classical computational complexity of CSPs. Despite the provable existence of intermediate (say, between P and NP-complete, assuming $P \neq NP$) problems, research in computational complexity has produced a widely known informal thesis that “natural problems are almost always complete for standard complexity classes”. CSPs have been actively used to support and refine this thesis. More precisely, several restricted forms of CSP have been investigated in depth. One of the main types of restrictions is the *constraint language* restriction, i.e., a restriction on the available types of constraints. By choosing an appropriate constraint language, one can obtain many well-known computational problems from graph theory, logic, and algebra. The study of the constraint language restriction is driven by the CSP *Dichotomy Conjecture* of Feder and Vardi which states that, for each fixed constraint language, the corresponding CSP is either in P or NP-complete. There are similar dichotomy conjectures concerning other complexity classes (e.g. L and NL). Recent breakthroughs in the complexity of CSP have been made possible by the introduction of the universal-algebraic approach, which extracts algebraic structure from the constraint language and uses it to analyse problem instances. The above conjectures have algebraic versions which also predict in algebraic terms where the boundary between harder problems and easier problems lies. The algebraic approach has been applied to prove the Dichotomy Conjecture in many important special cases (e.g. Bulatov’s dichotomy theorems for 3-valued and conservative CSPs), but the general problem remains open. Barto and Willard described the current state-of-the-art in proving this conjecture, gave insights into the main stumbling blocks (notably, the convoluted ways in which systems of linear equations appear in constraint problems), and outlined avenues of attack on those obstacles. Kozik gave a new simplified algorithm for CSPs solvable by local consistency methods, confirming an earlier conjecture. Brown-Cohen presented new results leading to closer interchange of ideas between algebraic and probabilistic approaches to CSPs.

Valued CSP is a significant generalisation of CSP that involves both feasibility and optimisation aspects. The complexity of language-based restriction for VCSPs was considered in the talks by Kolmogorov, Thapper, and Živný. Very strong result in this direction were reported, especially the full description of tractable cases modulo CSP, which closes a sequence of strong and unexpected results on VCSPs obtained during last five years.

The complexity of counting solutions for CSPs, with many results, was investigated by Goldberg, Jerrum, and Richerby.

Along with the constraint language restriction on CSP, the other main type is the structural restriction (i.e. restriction on the immediate interaction between variables in instances). Structural restrictions leading to tractability are well-understood, by results of Grohe and Marx. The so-called “hybrid” tractability in CSP, which is tractability that cannot be attributed to a constraint language restriction or to a structural restriction alone, has not received a great deal of attention yet, and is one of the possible avenues of future work. Rolínek, Scarcello, and Živný described recent results on hybrid tractability for CSPs and VCSPs, including counting problems.

Approximability of CSPs. The use of approximation algorithms is one of the most fruitful approaches to coping with NP-hardness. Hard optimization problems, however, exhibit different behavior with respect to approximability, making it an exciting, and by now, well-developed but far from fully understood, research area. The CSP has always played an important role in the study of approximability. For example, it is well known that the famous PCP theorem has an equivalent reformulation in terms of inapproximability of a

certain CSP; moreover, the recent combinatorial proof of this theorem by Dinur in 2006 deals entirely with CSPs. The first optimal inapproximability results by Håstad in 2001 were about certain CSPs, and they led to the study of a new hardness notion called *approximation resistance* (which, intuitively, means that a problem cannot be approximated beyond the approximation ratio given by picking an assignment uniformly at random, even on almost satisfiable instances). Many CSPs have been classified as to whether they are approximation resistant but there is not even a reasonable conjecture for a full classification. Lee and Tulsiani presented new results on approximation resistance.

Many approximation algorithms for CSPs are based on the Sum-of-Squares method, Linear Programming and Semidefinite Programming. Recent developments in proving lower bounds for such algorithms were presented by Chan and Steurer.

Improved approximation algorithms for certain infinite-domain CSPs related to correlation clustering were given by K. Makarychev.

New applications of algebraic approach to investigate approximability of CSPs were given by Austrin and Dalmau.

Parameterized complexity of CSPs. A different way to cope with NP-hardness is provided by parameterized complexity, which relaxes the notion of tractability as polynomial-time solvability to allow non-polynomial dependence on certain problem-specific parameters. A whole new set of interesting questions arises if we look at CSPs from this point of view. Most CSP dichotomy questions can be revisited by defining a parameterized version; so far, very little work was done in this direction compared to investigations in classical complexity. A new research direction (often called “parameterizing above the guaranteed bound”) led to unexpected positive results for Max r -SAT by Alon *et al.* in 2010. In this direction, the basic question is to decide the fixed-parameter tractability of the following type of problems: if some easily computable estimate guarantees satisfaction at least E constraints, find an assignment that satisfies at least $E + k$ constraints. Y. Makarychev presented recent results, including approximation issues, in this direction that concern the so-called ordering CSP. Wahlström and Yoshida described how algorithms for this problem, also for VCSP, can be designed when the estimate is given by the Linear Programming relaxation.

Logic and the complexity of CSP. Starting from earlier work by Kolaitis and Vardi, concepts and techniques from logic have provided unifying explanations for many tractable CSPs. This has led to the pursuit of classifications of CSP with respect to *descriptive complexity*, i.e. definability in a given logic. Logics considered in this context include first order logic and its extensions, finite-variable logics, the logic programming language Datalog and its fragments. Kazda presented his recent results on the two most important open problems on descriptive complexity of CSPs, where he showed that one of these problems reduces to the other. These results are also related to dichotomy questions for complexity classes L and NL.

The CSP can be recast as the problem of deciding satisfiability of existential conjunctive formulas. Chen described recent results in this direction that also involve counting and parameterised complexity. Natural extension of this framework that also allows universal quantifiers is known as the Quantified CSP (QCSP). New results on the complexity of language-restricted QCSPs were presented by Martin. Zhuk gave a proof of an algebraic result that has direct strong consequences for complexity classification of QCSPs.

Bodirsky and Pinsker presented latest developments in infinite domain CSPs, obtained via a mixture of model-theoretic and algebraic methods.

Ochremiak investigated finite-domain CSPs on infinite instances definable by formulas in first-order logic.

2 Table of Contents

Executive Summary

Andrei A. Bulatov, Venkatesan Guruswami, Andrei Krokhin, and Dániel Marx . . . 22

Overview of Talks


(2+ ε)-SAT is NP-hard	
<i>Per Austrin</i>	28
CSPs over hereditarily semisimple algebras	
<i>Libor Barto</i>	28
Max-Closed Semilinear Constraints	
<i>Manuel Bodirsky</i>	28
Correlation Decay from Cyclic Polymorphisms	
<i>Jonah Brown-Cohen</i>	29
Sum of Squares Lower Bounds from Pairwise Independence	
<i>Siu On Chan</i>	29
The Parameterized Complexity Classification of $\#P$ and the Logic of $\#P$ -Counting	
Answers to Existential Positive Queries	
<i>Hubie Chen</i>	29
Approximating Bounded-Degree Boolean Counting CSPs	
<i>Leslie Ann Goldberg</i>	30
Approximately counting H-colourings is $\#BIS$ -hard	
<i>Mark R. Jerrum</i>	31
Linear Datalog and n -permutability implies symmetric Datalog	
<i>Alexandr Kazda</i>	32
The Complexity of General-Valued CSPs	
<i>Vladimir Kolmogorov</i>	32
Congruence join semi-distributive varieties and constraint satisfaction problems	
<i>Marcin Kozik</i>	32
Towards a Characterization of Approximation Resistance for Symmetric CSPs	
<i>Euiwoong Lee</i>	33
Correlation Clustering on Complete Graphs	
<i>Konstantin Makarychev</i>	33
Satisfiability of Ordering CSPs Above Average Is Fixed-Parameter Tractable	
<i>Yury Makarychev</i>	34
Algebra and the Complexity of Quantified Constraints	
<i>Barnaby Martin</i>	34
Deciding FO-definable CSP instances	
<i>Joanna Ochremiak</i>	34
Constraint Satisfaction Problems on K_n -free graphs	
<i>Michael Pinsker</i>	35
Counting Matrix Partitions of Graphs	
<i>David Richerby</i>	35

Effectiveness of structural restrictions for hybrid CSPs	
<i>Michal Rolínek</i>	36
Hybrid Tractability of the Counting Problem, with a case study from Game Theory	
<i>Francesco Scarcello</i>	36
Lower bounds for LP/SDP formulations of CSPs	
<i>David Steurer</i>	37
The power of Sherali-Adams relaxations for valued CSPs	
<i>Johan Thapper</i>	37
A Characterization of Strong Approximation Resistance	
<i>Madhur Tulsiani</i>	37
Parameterized VCSPs, Euler digraphs and diamond lattices	
<i>Magnus Wahlström</i>	38
Maltsev constraints revisited	
<i>Ross Willard</i>	38
Half-Integrality, LP-Branching and FPT Algorithms	
<i>Yuichi Yoshida</i>	39
EGP/PGP dichotomy	
<i>Dmitriy Zhuk</i>	39
Hybrid (V)CSPs	
<i>Stanislav Živný</i>	40
Participants	41

3 Overview of Talks

3.1 $(2+\epsilon)$ -SAT is NP-hard

Per Austrin (KTH Royal Institute of Technology, SE)

License  Creative Commons BY 3.0 Unported license
© Per Austrin

Joint work of Austrin, Per; Guruswami, Venkatesan; Håstad, Johan

Main reference P. Austrin, J. Håstad, V. Guruswami, “ $(2 + \epsilon)$ -Sat Is NP-Hard,” in Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS’14), pp. 1–10, IEEE 2014.

URL <http://dx.doi.org/10.1109/FOCS.2014.9>

We prove the following hardness result for a natural promise variant of the SAT problem: given a CNF formula where each clause has width w and the guarantee that there exists an assignment satisfying at least $g = w/2 - 1$ literals in each clause, it is NP-hard to find a satisfying assignment to the formula (that sets at least one literal to true in each clause). On the other hand, when $g = w/2$, it is easy to find a satisfying assignment via simple generalizations of the algorithms for 2-SAT.

We also generalize this to prove strong NP-hardness for discrepancy problems with small size sets.

3.2 CSPs over hereditarily semisimple algebras

Libor Barto (Charles University – Prague, CZ)

License  Creative Commons BY 3.0 Unported license
© Libor Barto

In our latest joint effort with Marcin Kozik to resolve the CSP tractability conjecture we considered an inductive strategy to solve the problem. Most of the facts needed to make this strategy work turned out to be incorrect. Luckily, at least the base induction step works. This gives, for example, the following: $CSP(A)$ is tractable, whenever A is a Taylor algebra (ie, satisfies the necessary condition for tractability) and each subalgebra of A is semisimple.

3.3 Max-Closed Semilinear Constraints

Manuel Bodirsky (TU Dresden, DE)

License  Creative Commons BY 3.0 Unported license
© Manuel Bodirsky

A subset of Q^n is called semilinear if it can be defined by Boolean combinations of linear inequalities. Such a subset is called tropically convex if it is preserved by taking componentwise maximum and translations. We show that the Constraint Satisfaction Problem for tropically convex semilinear relations is in NP intersected coNP. This was previously only known for the substantially smaller class of max-atoms constraints.

3.4 Correlation Decay from Cyclic Polymorphisms

Jonah Brown-Cohen (University of California – Berkeley, US)

License © Creative Commons BY 3.0 Unported license
© Jonah Brown-Cohen

Joint work of Brown-Cohen, Jonah; Raghavendra, Prasad

In recent work with Prasad Raghavendra we prove that cyclic polymorphisms exhibit a correlation decay phenomenon. In particular, suppose D is a distribution on satisfying assignments to some constraint. If D has no perfect correlations between variables, then repeatedly applying a cyclic polymorphism to D will decay correlation until D becomes a product distribution. I will explain how this correlation decay theorem is proved, and give some toy examples of possible applications to algorithms for CSPs which admit cyclic polymorphisms.

3.5 Sum of Squares Lower Bounds from Pairwise Independence

Siu On Chan (The Chinese University of Hong Kong, HK)

License © Creative Commons BY 3.0 Unported license
© Siu On Chan

Joint work of Barak, Boaz; Chan, Siu On; Kothari, Pravesh

Main reference B. Barak, S. O. Chan, P. Kothari, “Sum of Squares Lower Bounds from Pairwise Independence,” arXiv:1501.00734v2 [cs.CC], 2015.

URL <http://arxiv.org/abs/1501.00734v2>

We prove that for every $\varepsilon > 0$ and k -ary predicate P that supports a pairwise independent distribution, there exists an instance I of the MaxP constraint satisfaction problem on n variables such that no assignment can satisfy more than a $|P^{-1}(1)|/2^k + \varepsilon$ fraction of I ’s constraints but the degree $\Omega(n)$ Sum of Squares semidefinite programming hierarchy cannot certify that I is unsatisfiable. Similar results were previously only known for weaker hierarchies.

3.6 The Parameterized Complexity Classification of – and the Logic of – Counting Answers to Existential Positive Queries

Hubie Chen (Universidad del País Vasco – Donostia, ES)

License © Creative Commons BY 3.0 Unported license
© Hubie Chen

Joint work of Chen, Hubie; Mengel, Stefan

We consider the computational complexity of the problem of counting the number of answers to a logical formula on a finite structure.

We present two contributions.

First, in the setting of parameterized complexity, we present a classification theorem on classes of existential positive queries. In particular, we prove that (relative to the problem at hand) any class of existential positive formulas is interreducible with a class of primitive positive formulas. In the setting of bounded arity, this allows us to derive a trichotomy theorem indicating the complexity of any class of existential positive formulas, as we previously proved a trichotomy theorem on classes of primitive positive formulas (Chen

and Mengel '15). This new trichotomy theorem generalizes and unifies a number of existing classification results in the literature, including the classifications on model checking primitive positive formulas (Grohe '07), model checking existential positive formulas (Chen '14), and counting homomorphisms (Dalmau and Jonsson '04).

Our second contribution is to introduce and study an extension of first-order logic in which algorithms for the counting problem at hand can be naturally and conveniently expressed. In particular, we introduce a logic which we call $\#$ -logic where the evaluation of a so-called $\#$ -sentence on a structure yields a natural number, as opposed to just a propositional value (true or false) as in usual first-order logic. We discuss the width of a formula as a natural complexity measure and show that this measure is meaningful in $\#$ -logic and that there is an algorithm that minimizes width in the “existential positive fragment” of $\#$ -logic.

This is joint work with Stefan Mengel.

3.7 Approximating Bounded-Degree Boolean Counting CSPs

Leslie Ann Goldberg (University of Oxford, GB)

License  Creative Commons BY 3.0 Unported license
© Leslie Ann Goldberg

Joint work of Galanis, Andreas; Goldberg, Leslie Ann

Main reference A. Galanis, L. A. Goldberg, “The complexity of approximately counting in 2-spin systems on k -uniform bounded-degree hypergraphs,” arXiv:1505.06146v3 [cs.CC], 2015.

URL <http://arxiv.org/abs/1505.06146v3>

The talk introduced some work on the complexity of approximately counting satisfying assignments of Boolean read- Δ CSPs (where each variable can be used at most Δ times). The work, which is joint with Andreas Galanis, is inspired by recent developments in the study of approximating partition functions. Here is an interesting connection which has been discovered recently by researchers in the area: for certain (weighted) Boolean binary constraints (called “anti-ferromagnetic 2-spin systems” in statistical physics) there are deep connections between the complexity of approximately counting satisfying assignments and a phenomenon known as “the uniqueness phase transition” which arises when the CSP is considered on the infinite Δ -regular tree.

I spent most of the talk explaining this connection. Our work is motivated by considering the extent to which the connection extends to a broader class of CSPs. We show that for every symmetric Boolean function f (apart from seven classes of trivial ones) there is a barrier Δ_0 such that for all $\Delta \geq \Delta_0$, it is NP-hard to approximate the partition function. The paper is available at <http://arxiv.org/abs/1505.06146>.

3.8 Approximately counting H -colourings is $\#BIS$ -hard

Mark R. Jerrum (Queen Mary University of London, GB)

License © Creative Commons BY 3.0 Unported license
© Mark R. Jerrum

Joint work of Galanis, Andreas; Goldberg, Leslie Ann; Jerrum, Mark

Main reference A. Galanis, L. A. Goldberg, M. Jerrum, “Approximately counting H -colourings is $\#BIS$ -Hard,” arXiv:1502.01335v1 [cs.CC], 2015.

URL <http://arxiv.org/abs/1502.01335v1>

We consider the problem of counting H -colourings, i.e., homomorphisms from an instance graph G to a fixed target graph H . In the CSP setting, this corresponds to the case of a constraint language with one symmetric binary relation. The complexity of computing an exact solution is well understood even in massively more general situations, thanks to a sequence of papers, starting with Dyer and Greenhill [3], and culminating (for the moment) with Cai and Chen [1]. The complexity of computing approximate solutions is much less well understood.

We show that for any fixed graph H without trivial components, it is as hard to approximate the number of H -colourings of a graph as it is to approximately count independent sets in a bipartite graph. The latter problem, called $\#BIS$, is a complete problem in an important complexity class for approximate counting, and is believed not to have an efficient approximation algorithm or FPRAS. If this is so, then our result shows that for every graph H without trivial components, the H -colouring counting problem has no FPRAS.

This problem was studied a decade ago by Goldberg, Kelk and Paterson [5]. They were able to show that approximately sampling H -colourings is $\#BIS$ -hard, but it was not known how to get the result for approximate counting. (For general H -colouring problems, there is a generic reduction from approximate counting to sampling due to Dyer, Goldberg and Jerrum [2], but still there is no known reduction in the opposite direction.)


The talk was based on the preprint [4], but the emphasis was on conveying the fundamentals of the complexity of approximate counting, and the flavour of the results, rather than on describing in detail the technicalities of the proof.

References

- 1 Jin-Yi Cai and Xi Chen. Complexity of counting CSP with complex weights. In *STOC'12 – Proceedings of the 2012 ACM Symposium on Theory of Computing*, pp. 909–919. ACM, New York, 2012.
- 2 M. E. Dyer, L. A. Goldberg, and M. Jerrum. Counting and sampling H -colourings. *Information and Computation*, 189(1):1–16, 2004.
- 3 M. E. Dyer and C. Greenhill. The complexity of counting graph homomorphisms. *Random Structures and Algorithms*, 17(3-4):260–289, 2000.
- 4 Andreas Galanis, Leslie Ann Goldberg and Mark Jerrum, Approximately counting H -colourings is $\#BIS$ -Hard. arXiv:1502.01335, February 2015.
- 5 L. A. Goldberg, S. Kelk, and M. Paterson. The complexity of choosing an H -coloring (nearly) uniformly at random. *SIAM J. Comput.*, 33(2):416–432, 2004.

3.9 Linear Datalog and n -permutability implies symmetric Datalog

Alexandr Kazda (IST Austria – Klosterneuburg, AT)

License  Creative Commons BY 3.0 Unported license
© Alexandr Kazda

The classification of CSPs that can be solved by the Datalog programming language and its fragments, linear and symmetric Datalog, is useful to understand the finer complexity of CSPs that lie in P. The full Datalog corresponds to CSPs solvable by local consistency, and evaluating linear resp. symmetric Datalog lies in NL resp. L.

In this talk, we show that if $CSP(A)$ is solvable by linear Datalog and A satisfies the additional algebraic condition of congruence n -permutability (for some n), then $CSP(A)$ is solvable by symmetric Datalog, which is the weakest of the fragments of Datalog.

3.10 The Complexity of General-Valued CSPs

Vladimir Kolmogorov (IST Austria – Klosterneuburg, AT)

License  Creative Commons BY 3.0 Unported license
© Vladimir Kolmogorov

Joint work of Kolmogorov, Vladimir; Krokhin, Andrei; Rolinek, Michal

Main reference V. Kolmogorov, A. Krokhin, M. Rolinek, “The Complexity of General-Valued CSPs,” arXiv:1502.07327v3 [cs.CC], 2015.

URL <http://arxiv.org/abs/1502.07327v3>

Consider a general valued language Γ . Kozik and Ochremiak recently showed that if the core of the language does not admit a cyclic fractional polymorphism of arity at least 2 then it is NP-hard. We prove that if this necessary condition is satisfied, and the underlying feasibility CSP is tractable, then Γ is tractable. The algorithm is a simple combination of the assumed algorithm for the feasibility CSP and the standard LP relaxation. As a corollary, we obtain that a dichotomy for ordinary CSPs would imply a dichotomy for general-valued CSPs.

Joint work with Andrei Krokhin and Michal Rolinek.

3.11 Congruence join semi-distributive varieties and constraint satisfaction problems

Marcin Kozik (Jagiellonian University – Kraków, PL)

License  Creative Commons BY 3.0 Unported license
© Marcin Kozik

I will present some new results on $SD(\vee)$ varieties: in particular a set of directed $SD(\vee)$ terms and an analogue of “graph absorbing” condition for congruence distributive varieties. I will discuss further applications of these results to solving CSPs for such algebras. Finally I will show examples of problematic algebras (in $SD(\vee)$) which present new problems while solving CSPs (compared to the CD case).

3.12 Towards a Characterization of Approximation Resistance for Symmetric CSPs

Euiwoong Lee (Carnegie Mellon University, US)

License © Creative Commons BY 3.0 Unported license
 © Euiwoong Lee
Joint work of Lee, Euiwoong; Guruswami, Venkatesan
Main reference V. Guruswami, E. Lee, “Towards a Characterization of Approximation Resistance for Symmetric CSPs,” ECCC, TR15-105, 2015.
URL <http://eccc.hpi-web.de/report/2015/105/>

A Boolean constraint satisfaction problem (CSP) is called approximation resistant if independently setting variables to 1 with some probability achieves the best possible approximation ratio for the fraction of constraints satisfied. We study approximation resistance of a natural subclass of CSPs that we call Symmetric Constraint Satisfaction Problems (SCSPs), where satisfaction of each constraint only depends on the number of true literals in its scope. Thus a SCSP of arity k can be described by a subset of allowed number of true literals.

For SCSPs without negation, we conjecture that a simple sufficient condition to be approximation resistant by Austrin and Hastad is indeed necessary. We show that this condition has a compact analytic representation in the case of symmetric CSPs (depending only on the gap between the largest and smallest numbers in S), and provide the rationale behind our conjecture. We prove two interesting special cases of the conjecture, (i) when S is an interval and (ii) when S is even. For SCSPs with negation, we prove that the analogous sufficient condition by Austrin and Mossel is necessary for the same two cases, though we do not pose an analogous conjecture in general.

3.13 Correlation Clustering on Complete Graphs

Konstantin Makarychev (Microsoft Corporation – Redmond, US)

License © Creative Commons BY 3.0 Unported license
 © Konstantin Makarychev


We give new rounding schemes for the standard linear programming relaxation of the correlation clustering problem (which can be seen as a constraint satisfaction problem), achieving approximation factors almost matching the integrality gaps:

- For complete graphs our approximation is $2.06 - \varepsilon$ for a fixed constant ε , which almost matches the previously known integrality gap of 2.
- For complete k -partite graphs our approximation is 3. We also show a matching integrality gap.
- For complete graphs with edge weights satisfying triangle inequalities and probability constraints, our approximation is 1.5, and we show an integrality gap of 1.2.

Our results improve a long line of work on approximation algorithms for correlation clustering in complete graphs, previously culminating in a ratio of 2.5 for the complete case by Ailon, Charikar and Newman (JACM’08). In the weighted complete case satisfying triangle inequalities and probability constraints, the same authors give a 2-approximation; for the bipartite case, Ailon, Avigdor-Elgrabli, Liberty and van Zuylen give a 4-approximation (SICOMP’12).

3.14 Satisfiability of Ordering CSPs Above Average Is Fixed-Parameter Tractable

Yury Makarychev (TTIC – Chicago, US)

License  Creative Commons BY 3.0 Unported license
© Yury Makarychev

We study the satisfiability of ordering constraint satisfaction problems (CSPs) above average. We prove the conjecture of Gutin, van Iersel, Mnich, and Yeo that the satisfiability above average of ordering CSPs of arity k is fixed-parameter tractable for every k . Previously, this was only known for $k = 2$ and $k = 3$. We also generalize this result to more general classes of CSPs, including CSPs with predicates defined by linear equations.

To obtain our results, we prove a new Bonami-type inequality for the Efron-Stein decomposition. The inequality applies to functions defined on arbitrary product probability spaces. In contrast to other variants of the Bonami Inequality, it does not depend on the mass of the smallest atom in the probability space. We believe that this inequality is of independent interest.

3.15 Algebra and the Complexity of Quantified Constraints

Barnaby Martin (Middlesex University, GB)

License  Creative Commons BY 3.0 Unported license
© Barnaby Martin

We survey connections between algebra and the complexity of Quantified Constraint Satisfaction Problems over finite templates. The complexity of such problems is well-known to hinge on their surjective polymorphisms, and early results studied the complexity divide between P and NP-hard. More recent work focuses on the gap between NP and PSPACE-hard and the connections to the so-called Polynomial Generate Powers property of the respective algebra's direct powers.

3.16 Deciding FO-definable CSP instances

Joanna Ochremiak (University of Warsaw, PL)

License  Creative Commons BY 3.0 Unported license
© Joanna Ochremiak

Joint work of Klin, Bartek; Kopczynski, Eryk; Ochremiak, Joanna; Toruńczyk, Szymon
Main reference B. Klin, E. Kopczynski, J. Ochremiak, S. Toruńczyk, “Locally Finite Constraint Satisfaction Problems,” in Proc. of the 30th Annual ACM/IEEE Symp. on Logic in Computer Science (LICS’15), pp. 475–486, IEEE, 2015; pre-print available from author’s webpage.
URL <http://dx.doi.org/10.1109/LICS.2015.51>
URL <http://www.mimuw.edu.pl/~ochremiak/papers/lics15.pdf>

In this talk I showed how to decide infinite CSP instances that exhibit high level of symmetry. I considered instances founded upon a fixed infinite relational structure, and defined by finitely many FO formulas. The number of elements and constraints in such instances is usually infinite, but thanks to the finite presentation they can be treated as an input for algorithms. The decidability proof is based on results in topological dynamics. Moreover, I showed tight complexity bounds for this problem.

3.17 Constraint Satisfaction Problems on K_n -free graphs

Michael Pinsker (*University Paris-Diderot, FR*)

License © Creative Commons BY 3.0 Unported license
© Michael Pinsker

We present a dichotomy result for constraint satisfaction problems over K_n -free graphs, i.e., graphs which do not contain any clique of size n , where $n \geq 3$ is a fixed natural number. In these problems, the input consists of variables and constraints about them which have to be taken from a fixed finite set of quantifier-free formulas in the language of graphs. The decision problem is whether the variables can be assigned vertices in a K_n -free graph so that all constraints are satisfied.

Using methods from infinite constraint satisfaction, in particular Ramsey theory, we show that all such problems are either in **P** or **NP**-complete.

This result is a (incomparable) variant of Schaefer's theorem for graphs proven by Bodirsky + Pinsker in 2011, and we will compare the proof of that theorem with the one of the theorem we present.

3.18 Counting Matrix Partitions of Graphs

David Richerby (*Oxford University, GB*)

Joint work of Dyer, Martin; Goebel, Andreas; Goldberg, Leslie Ann; McQuillan, Colin; Yamakami, Tomoyuki
License © Creative Commons BY 3.0 Unported license
© David Richerby

A matrix partition of a graph G is a partition of its vertices into parts V_1, \dots, V_k whose connections to one another are specified by a symmetric k -by- k $\{0, 1, \star\}$ matrix M according to the following rules:

- if $M_{i,j} = 1$, there must be an edge in G between every pair of distinct vertices $x \in V_i$ and $y \in V_j$;
- if $M_{i,j} = 0$, there must be no edge in G between any vertex in V_i and any in V_j
- if $M_{i,j} = \star$, there may be any pattern of edges and non-edges between V_i and V_j .

If M contains no 1s, M -partitions are equivalent to graph homomorphisms. General matrix problems can also be formulated as CSPs with restrictions on the input.

Matrix partitions naturally encode many classes of graphs, such as split graphs and (a, b) -graphs and graph structures such as k -colourings, skew cutsets and various generalizations of clique-cross partitions.

I will discuss the complexity of counting matrix partitions of graphs. This is joint work with Martin Dyer, Andreas Goebel, Leslie Ann Goldberg, Colin McQuillan and Tomoyuki Yamakami.

3.19 Effectiveness of structural restrictions for hybrid CSPs

Michal Rolinek (*IST Austria – Klosterneuburg, AT*)

License © Creative Commons BY 3.0 Unported license
© Michal Rolinek

Joint work of Kolmogorov, Vladimir; Rolinek, Michal; Takhanov, Rustem

Main reference V. Kolmogorov, M. Rolinek, R. Takhanov, “Effectiveness of Structural Restrictions for Hybrid CSPs,” arXiv:1504.07067v3 [cs.CC], 2015.

URL <http://arxiv.org/abs/1504.07067v3>

We focus on the hybrid setting of the CSP that restricts both sides of the homomorphism problem simultaneously. It assumes that the input belongs to a certain class of relational structures (called a structural restriction in this paper). We study which structural restrictions are effective, i.e. there exists a fixed template (from a certain class of languages) for which the problem is tractable when the input is restricted, and NP-hard otherwise. We provide a characterization for structural restrictions that are closed under inverse homomorphisms and show its implications (for example for minor-closed families, or ordered CSP). We extend the results to certain Valued CSPs (namely conservative valued languages).

3.20 Hybrid Tractability of the Counting Problem, with a case study from Game Theory

Francesco Scarcello (*University of Calabria, IT*)

License © Creative Commons BY 3.0 Unported license
© Francesco Scarcello

Joint work of Scarcello, Francesco; Greco, Gianluigi

Main reference F. Scarcello, G. Greco, “Counting solutions to conjunctive queries: structural and hybrid tractability,” in Proc. of the 33rd ACM SIGMOD/SIGACT/SIGART Symp. on Principles of Database Systems, pp. 132–143, ACM, 2014.

URL <http://dx.doi.org/10.1145/2594538.2594559>

Counting the number of solutions of CSPs with output variables is an intractable problem, formally #P-hard, even over classes of acyclic instances. We describe structural methods that allow us to identify large islands of tractability, such as the method based on #-generalized hypertree decompositions.

Based on this notion, a “hybrid” decomposition method is eventually conceived, where structural properties of the left-hand structure (i.e., properties of constraint scopes and output variables) are exploited in combination with properties of the right-hand structure (i.e., the values). Intuitively, such features may induce different structural properties that are not identified by the “worst-possible” perspective of purely structural methods.

We eventually describe a successful application of such CSP techniques to solve a game theory problem, namely, the polynomial-time computation of the Shapley value in allocations games. These are coalitional games defined in the literature as a way to analyze fair division problems of indivisible goods.

References

- 1 G. Greco, F. Lupia, and F. Scarcello, *Structural Tractability of Shapley and Banzhaf Values in Allocation Games*, in Proceedings of IJCAI 2015, pp. 547–553, 2015.
- 2 G. Greco and F. Scarcello, *Mechanisms for Fair Allocation Problems: No-Punishment Payment Rules in Verifiable Settings*. J. Artif. Intell. Res. (JAIR) 49:403–449 (2014).
- 3 G. Greco and F. Scarcello, *Counting Solutions to Conjunctive Queries: Structural and Hybrid Tractability*, in Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS’14), ACM, 2014.

3.21 Lower bounds for LP/SDP formulations of CSPs

David Steurer (Cornell University – Ithaca, US)

License © Creative Commons BY 3.0 Unported license
© David Steurer

We introduce a method for proving lower bounds on the efficacy of semidefinite programming (SDP) relaxations for combinatorial problems.

3.22 The power of Sherali-Adams relaxations for valued CSPs

Johan Thapper (University Paris-Est – Marne-la-Vallée, FR)

License © Creative Commons BY 3.0 Unported license
© Johan Thapper
Joint work of Thapper, Johan; Živný, Stanislav

We give an algebraic characterisation of when some bounded level of the Sherali-Adams hierarchy solves the valued constraint satisfaction problem (VCSP) parameterised by a valued constraint language to optimality. We also give examples of classes of VCSPs that are NP-hard unless they are solved by such an LP relaxation. Our result is fundamentally based on the algebraic characterisation of the notion of bounded relational width for ordinary decision CSPs by Larose/Zádori (2007) and Barto/Kozik (2014).

This is joint work with Stanislav Živný.

3.23 A Characterization of Strong Approximation Resistance

Madhur Tulsiani (TTIC – Chicago, US)

License © Creative Commons BY 3.0 Unported license
© Madhur Tulsiani
Joint work of Jones, Mark; Sheng, Bin

For a predicate $f : \{-1, 1\}^k \rightarrow \{0, 1\}$ with $E[f] = \rho$, we call the predicate strongly approximation resistant if given a near-satisfiable instance of $CSP(f)$, it is computationally hard to find an assignment such that the fraction of constraints satisfied is outside the range $(\rho - \varepsilon, \rho + \varepsilon)$ for every $\varepsilon > 0$.

We present a characterization of strongly approximation resistant predicates under the Unique Games Conjecture. We also present characterizations in the mixed linear and semidefinite programming hierarchy and the Sherali-Adams linear programming hierarchy. In the former case, the characterization coincides with the one based on UGC. Each of the two characterizations is in terms of existence of a probability measure on a natural convex polytope associated with the predicate.

The predicate is called approximation resistant if given a near-satisfiable instance of $CSP(f)$, it is computationally hard to find an assignment such that the fraction of constraints satisfied is at least $\rho + \varepsilon$. When the predicate is odd, i.e. $f(-z) = 1 - f(z)$ for all z in $\{-1, 1\}^k$, it is easily observed that the notion of approximation resistance coincides with that of strong approximation resistance. Hence for odd predicates, in all the above settings, our characterization of strong approximation resistance is also a characterization of approximation resistance.

3.24 Parameterized VCSPs, Euler digraphs and diamond lattices

Magnus Wahlström (Royal Holloway University of London, GB)

License  Creative Commons BY 3.0 Unported license
© Magnus Wahlström

Joint work of Jones, Mark; Sheng, Bin; Wahlström, Magnus

VCSPs (Valued CSPs, an optimisation variant of CSP) is a general problem framework that encompasses many problems of independent interest. In particular, many of the success stories of parameterized complexity, and many of its still open problems, can be phrased as questions about the parameterized complexity properties (in particular, FPT algorithms) of particular VCSPs.

P vs NP dichotomies for VCSP are known (Thapper and Živný, FOCS 2012/STOC 2013, extended by Kolmogorov, Krokhin, and Rolinek, 2015), showing essentially that the P-time solvable cases coincide with a natural LP-relaxation of the problem. These results, and the LP-relaxation, have also shown to be useful for FPT algorithms – the LP-branching approach has been shown to imply some very powerful FPT algorithms (Wahlström, SODA 2014/Iwata, Wahlström, and Yoshida, 2014), and the results of Thapper and Živný are central to the correctness of these algorithms. More concretely, it was shown that so-called k -submodular functions, which were shown to be tractable by Thapper and Živný, have direct applications for VCSP FPT algorithms.

In this talk, we survey the above results and discuss an attempt at handling the directed versions of these problems. We show how certain digraph cut problems (in particular, directed multiway cut) can be recast as optimisation of functions which are submodular over a diamond lattice. Although we cannot yet recreate the existing FPT results in this way, we get the following results:

- A 2-approximation for directed multiway cut (previously shown by Naor and Zosin, SICOMP 2001);
- A polynomial-time algorithm for directed multiway cut in Euler digraphs (the existence of such an algorithm was obscure, but implied by results of Frank, 1989).

Thanks to the algebraic VCSP setting, the results also extend to fairly natural variations of problems where arcs are labelled by bijections over a set of labels (e.g., directed versions of the so-called unique label cover and group feedback arc set problems).

Joint work with Mark Jones and Bin Sheng.

3.25 Maltsev constraints revisited

Ross Willard (University of Waterloo, CA)

License  Creative Commons BY 3.0 Unported license
© Ross Willard

Once upon a time, the (finite) CSP Dichotomy Conjecture was a Big Deal. However, since 2009 there has been no significant progress in solving it. Why? Because we are stupid, and the conjecture is hard, and most sensible researchers have moved on, but mostly because we lack algorithms.

Only two general algorithms for finite CSPs are known:

1. local consistency checking and
2. a generalization, largely due to Bulatov and Dalmau, of the Feder-Vardi algorithm for subgroup CSPs.

Each algorithm has its own naturally defined scope, and while there was early hope on the part of algebraists that some combination of the two algorithms would magically solve the Dichotomy Conjecture, such a solution hasn't materialized.

The problem, as I see it, is that the Bulatov-Dalmau algorithm is too simple. Like the Feder-Vardi algorithm it generalizes, the Bulatov-Dalmau algorithm neatly finesses the need to analyze the convoluted ways by which constraint systems can encode linear equations. It is now time, I suggest, to undertake this analysis. In this lecture I will sketch my recent (so far unsuccessful) efforts to fully understand encoded linear systems in constraint systems over finite templates having a Maltsev polymorphism.

3.26 Half-Integrality, LP-Branching and FPT Algorithms

Yuichi Yoshida (National Institute of Informatics – Tokyo, JP)

License © Creative Commons BY 3.0 Unported license
© Yuichi Yoshida

Joint work of Yoichi Iwata; Magnus Wahlström; Yoshida, Yuichi

A recent trend in parameterized algorithms is the application of polytope tools to FPT algorithms (e.g., Cygan et al., 2011; Narayanaswamy et al., 2012). Although this approach has yielded significant speedups for a range of important problems, it requires the underlying polytope to have very restrictive properties, including half-integrality and Nemhauser-Trotter-style persistence properties. To date, these properties are essentially known to hold only for two classes of polytopes, covering the cases of Vertex Cover (Nemhauser and Trotter, 1975) and Node Multiway Cut (Garg et al., 1994).

Taking a slightly different approach, we view half-integrality as a discrete relaxation of a problem, e.g., a relaxation of the search space from $\{0, 1\}^V$ to $\{0, 1/2, 1\}^V$ such that the new problem admits a polynomial-time exact solution. Using tools from CSP (in particular Thapper and Živný, 2012) to study the existence of such relaxations, we are able to provide a much broader class of half-integral polytopes with the required properties.

Our results unify and significantly extend the previously known cases, and yield a range of new and improved FPT algorithms, including an $O^*(|\Sigma|^{2k})$ -time algorithm for node-deletion Unique Label Cover and an $O^*(4k)$ -time algorithm for Group Feedback Vertex Set where the group is given by oracle access. The latter result also implies the first single-exponential time FPT algorithm for Subset Feedback Vertex Set, answering an open question of Cygan et al. (2012). Additionally, we propose a network-flow-based approach to solve several cases of the relaxation problem. This gives the first linear-time FPT algorithm to edge-deletion Unique Label Cover.

Joint work with Yoichi Iwata and Magnus Wahlström.

3.27 EGP/PGP dichotomy


Dmitriy Zhuk (Moscow State University, RU)

License © Creative Commons BY 3.0 Unported license
© Dmitriy Zhuk

For each algebra A , we count the minimal number $g(n)$ of generators of A^n as a function of n . We show that if $g(n)$ is not bounded from above by a polynomial (PGP) then $g(n)$ is bounded from below by an exponential function (EGP).

3.28 Hybrid (V)CSPs

Stanislav Živný (Oxford University, GB)

License  Creative Commons BY 3.0 Unported license
© Stanislav Živný

Given two classes A and B of finite relational structures over the same signature, the constraint satisfaction problem $CSP(A, B)$ amounts to, given structures $I \in A$ and $J \in B$, deciding whether there is a homomorphism from I to J . While the computational complexity of $CSP(A, -)$ is well understood (here $-$ denotes the class of all structures), the computational complexity of $CSP(-, B)$ is still open even if B consists of a single structure; these non-uniform CSPs have attracted a lot of attention due to the success of the so-called algebraic approach.

In this talk I will survey known results (older and recent ones) on so-called hybrid CSPs, which restrict both A and B . I will also mention known results for hybrid Valued CSPs.

Participants

- Albert Atserias
UPC – Barcelona, ES
- Per Austrin
KTH Royal Institute of
Technology, SE
- Libor Barto
Charles University – Prague, CZ
- Manuel Bodirsky
TU Dresden, DE
- Jonah Brown-Cohen
University of California –
Berkeley, US
- Andrei A. Bulatov
Simon Fraser University –
Burnaby, CA
- Catarina Alexandra Carvalho
University of Hertfordshire, GB
- Siu On Chan
The Chinese University of Hong
Kong, HK
- Hubie Chen
Universidad del País Vasco –
Donostia, ES
- Victor Dalmau
UPF – Barcelona, ES
- Laszlo Egri
Simon Fraser University –
Burnaby, CA
- Serge Gaspers
UNSW – Sydney, AU
- Leslie Ann Goldberg
University of Oxford, GB
- Venkatesan Guruswami
Carnegie Mellon University, US
- Mark R. Jerrum
Queen Mary University of
London, GB
- Peter Jonsson
Linköping University, SE
- Alexandr Kazda
IST Austria –
Klosterneuburg, AT
- Vladimir Kolmogorov
IST Austria –
Klosterneuburg, AT
- Marcin Kozik
Jagiellonian University –
Kraków, PL
- Andrei Krokhin
Durham University, GB
- Euiwoong Lee
Carnegie Mellon University, US
- Konstantin Makarychev
Microsoft Corporation –
Redmond, US
- Yury Makarychev
TTIC – Chicago, US
- Rajsekar Manokaran
Indian Institute of Technology –
Madras, IN
- Barnaby Martin
Middlesex University, GB
- Dániel Marx
Hungarian Academy of Sciences –
Budapest, HU
- Neeldhara Misra
Indian Institute of Science, IN
- Joanna Ochremiak
University of Warsaw, PL
- Michael Pinski
University Paris-Diderot, FR
- David Richerby
Oxford University, GB
- Michal Rolinek
IST Austria –
Klosterneuburg, AT
- Francesco Scarcello
University of Calabria, IT
- David Steurer
Cornell University – Ithaca, US
- Stefan Szeider
TU Wien, AT
- Johan Thapper
University Paris-Est –
Marne-la-Vallée, FR
- Madhur Tulsiani
TTIC – Chicago, US
- Matt Valeriot
McMaster University –
Hamilton, CA
- Magnus Wahlström
Royal Holloway University of
London, GB
- Ross Willard
University of Waterloo, CA
- Yuichi Yoshida
National Institute of Informatics –
Tokyo, JP
- Dmitriy Zhuk
Moscow State University, RU
- Stanislav Zivny
Oxford University, GB



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 Leibniz Institute for Psychology Information – 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.

Perspectives Workshop July 19–24, 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.5.7.42

Edited in cooperation with Christoph Lange

1 Executive Summary

Alexander Garcia Castro

Janna Hastings

Christoph Lange

Robert Stevens

Erich Weichselgartner

License © Creative Commons BY 3.0 Unported license

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

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, its tools and 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 culture. An impressive digital infrastructure has been built; this allows us to correlate



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

Digital Scholarship and Open Science in Psychology and the Behavioral Sciences, *Dagstuhl Reports*, Vol. 5, Issue 7, pp. 42–68

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



Dagstuhl Reports

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

information from genomes to diseases, and, by doing so, to support movements such as panomic studies and 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 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

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

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 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>Alexander Garcia Castro, Janna Hastings, Christoph Lange, Robert Stevens, and Erich Weichselgartner</i>	42
--	----

Overview of Talks

Mental illnesses, knowledge representation and data sharing <i>Xavier Aime</i>	47
The Human Behaviour Project <i>Dietrich Albert</i>	47
Data Archiving and Sharing Confidential Data <i>George Alter</i>	48
#dsos requires a digital infrastructure <i>Bjoern Brembs</i>	49
Research Objects for improved sharing and reproducibility in Psychology and Behavioural Sciences <i>Oscar Corcho</i>	49
Estimating the Reproducibility of Psychological Science <i>Susann Fiedler</i>	50
Goals of the Seminar on Digital Scholarship and Open Science in Psychology and the Behavioral Sciences <i>Alexander Garcia Castro</i>	50
‘Don’t Publish, Release’ Revisited <i>Paul Groth</i>	51
Open Science Lessons Learned at Mendeley <i>William Gunn</i>	51
The role of standards and ontologies in tackling reproducibility <i>Janna Hastings</i>	52
Developing reproducible and reusable methods through research software engineering <i>Caroline Jay</i>	53
Open science, mega analyses and problems in understanding the genetics of psychiatric disorders <i>Iris-Tatjana Kolassa</i>	54
Advancing Psychology and Behavioral Sciences in Brazil and World Wide <i>Silvia H. Koller</i>	55
Scholarly Communication and Semantic Publishing: Technical Challenges, and Recent Applications to Social Sciences <i>Christoph Lange</i>	56
Defining the Scholarly Commons: Are We There Yet: Summary of my presentation and some thoughts on the workshop <i>Maryann Martone</i>	59

Cognitive ontologies, data sharing, and reproducibility	
<i>Russell Poldrack</i>	61
Open Journal Systems: Introduction, Preview, and Community	
<i>Alec Smecher</i>	62
Principles, Programs and Pilots for Open Science and Digital Scholarship at Elsevier	
<i>Daniel Staemmler</i>	62
Open data and the need for ontologies	
<i>Robert Stevens</i>	64
Infrastructural Services for the Scientific Community provided by the American Psychological Association	
<i>Gary VandenBos</i>	65
Hijacking ORCID	
<i>Hal Warren</i>	66
PsychOpen – The European Open-Access Publishing Platform for Psychology	
<i>Erich Weichselgartner</i>	66
Participants	68

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

URL <http://europepmc.org/abstract/MED/25991135>

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 B. Brembs, K. Button, M. Munafò, “Deep impact: unintended consequences of journal rank,” *Frontiers in Human Neuroscience*, 7:291, 2013.

URL <http://dx.doi.org/10.3389/fnhum.2013.00291>

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 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

Main reference O. Ocho, “Research Objects for improved sharing and reproducibility,” Slides, 2015.

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.

3.6 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*, 349(6251), 2015.

URL <http://dx.doi.org/10.1126/science.aac4716>

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.

3.7 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


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.

3.8 ‘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, staring, 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.

3.9 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 content and built upon.

3.10 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

Tackling the reproducibility problem in research 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.

3.11 Developing reproducible and reusable methods through research software engineering

Caroline Jay (*University of Manchester, UK*)

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 IDInteraction 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.

3.12 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 \times 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 \times 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 \times 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 \times 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]). 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.

3.13 Advancing Psychology and Behavioral Sciences in Brazil and World Wide

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

License © Creative Commons BY 3.0 Unported license
 © Silvia H. Koller
URL <http://www.bvs-psi.org.br>

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.

3.14 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.

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).

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 disruption” 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.

- 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

⁴ <http://www.fiduswriter.org>

⁵ <https://pkp.sfu.ca/ojs/>

⁶ <http://www.gesis.org>

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, 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

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>

¹⁰ <http://lod.openaire.eu>

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.

3.15 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

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

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

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 ‘eScience 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.

3.16 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.

3.17 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.

3.18 Principles, Programs and Pilots for Open Science and Digital Scholarship at Elsevier

Daniel Staemmler (Elsevier Publishing – Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Daniel Staemmler

Elsevier supports the storing, sharing, discovering, and using of research data. Elsevier established a research data policy based on the STM Brussels Declaration 2007¹³ that “Raw research data should be made freely available to all researchers wherever possible” to help researchers to store, share, discover and use data.

The following principles underpin Elsevier’s data policy¹⁴:

- Research data should be made available free of charge to all researchers wherever possible and with minimal reuse restrictions.
- Researchers invest substantially to create and interpret data and others such as data archives, publishers, funders and institutions further add value and/or incur significant cost. In all such cases these contributions need to be recognized and valued.

¹³ http://www.stm-assoc.org/2007_11_01_Brussels_Declaration.pdf

¹⁴ <https://www.elsevier.com/about/company-information/policies/research-data>

- Expectations and practices around research data vary between disciplines and need to be taken into account.
- Platforms, publications, tools and services can enhance data by improving their discoverability, use, reuse, and citation.
- Standard identifiers, vocabularies, taxonomies, ontologies and entity resources enhance the discovery, management and use of data.

The following programs and pilots have been initiated:

1. *Data-linking program*¹⁵

Elsevier has an extensive program with 40+ leading domain-specific data repositories to interlink articles and data on ScienceDirect. This reciprocal linking aims to expand the availability of research data and improve the researcher workflow. Researchers – whether in the role of author or reader – benefit from both the increased discoverability of the data sets and seeing the data sets in the direct context of the research article. Linking through in-article accession numbers, data DOIs, or data banners are two examples on how this is being accomplished.

2. *Mendeley Data*

Allows researchers to store their research data online, so it can be cited and shared as well as securely saved in an online repository. DOIs and versioning of datasets, in compliance with Force11 standards, ensure that data citations are always valid. Mendeley Data is currently in beta phase.

3. *In-article data visualization*

a. iPlots¹⁶ – Displaying plot data in CSV format delivered by the author as supplementary material. Allows to access, explore, and download data behind plots.

b. 3D visualization tool¹⁷ – The goal is to enable Elsevier authors to showcase their 3D data, and to provide ScienceDirect users with a means to view and interact with these author-provided small to massive 3D datasets on a large number of devices with no additional plug-in required. These devices include smartphones, tablets, laptops and desktop computers.

4. *Open data and data profile*¹⁸

Increasing access to research data helps researchers to validate and build upon important discoveries and observations. With Elsevier's latest Open Data pilot we are providing authors with the opportunity to make their supplementary files with raw research data available open access on ScienceDirect.

5. *Data micro-articles*

Data journals, and data sections in existing journals, enable authors to have their research data peer-reviewed and cited. It will also make sure readers can find, use and analyze the data hosted in external databases or submitted as supplementary data. Examples of recently launched data journals are Genomics Data and Data in Brief.

6. *Standards bodies and working groups*

a. Joint Declaration of Data Citation Principles: best-practices to cite data in articles for better linking and credit

¹⁵ <http://www.elsevier.com/databaselinkin>

¹⁶ <https://www.elsevier.com/books-and-journals/content-innovation/iplots>


¹⁷ <http://www.elsevier.com/connect/bringing-3d-visualization-to-online-research-articles>

¹⁸ <https://www.elsevier.com/about/open-science/research-data/open-data>

- b. Research Data Alliance & ICSU World Data System: Tackling a broad range of interconnected issues around Data Publication (workflows, bibliometrics, cost recovery, services)
- 7. *Lay Summaries*
Making scientific research results accessible to the public by posting for each published article in the journal “Burnout Research” a lay summary explaining the main research findings. Burnout Research is one of Elsevier’s 270 open access titles and therefor freely available on the web (link to lay summaries).
- 8. *STM Digest*
¹⁹ STM Digest features lay summaries of science papers with societal impact. It is a collection of summaries of original research papers with social impact or a focus on policy. These summaries have the potential to make research more accessible, improve engagement in science, and benefit wider society. The initiative is a collaboration between Elsevier’s STM Journals group and the cloud-based research management and social collaboration platform, Mendeley.
- 9. *Atlas*²⁰
With over 1,800 journals publishing articles from across science, technology and health, Elsevier’s mission is to share some of the stories that matter. Each month Atlas showcases research that could significantly impact people’s lives around the world or has already done so. Bringing wider attention to this research will hopefully go some way to ensuring its successful implementation. Each month selecting a single article to be awarded “The Atlas” is facilitated by the Advisory Board. The winning research is presented in a lay-friendly, story format alongside interviews, expert opinions, and multimedia to reach a wide global audience.

3.19 Open data and the need for ontologies

Robert Stevens (University of Manchester, UK)

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

¹⁹ <http://www.elsevier.com/social-sciences/economics-and-finance/early-career-researchers>

²⁰ <https://www.elsevier.com/atlas/home>

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.

3.20 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 Association²¹ in Washington, DC, USA, and as the co-Editor of the Archives of Scientific Psychology²², 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 PsycINFO²³, a streaming Internet product. I also developed PsycARTICLES²⁴ (a full-text journal article database), PsycBOOKS²⁵ (a full-text book and book chapter database), PsycTESTS²⁶ (a measurement instrument database), and PsycTHERAPY²⁷ (a streaming video database of psychotherapy demonstrations). I am

²¹ <http://www.apa.org/>

²² <http://www.apa.org/pubs/journals/arc/>

²³ <http://www.apa.org/pubs/databases/psycinfo/>

²⁴ <http://www.apa.org/pubs/databases/psycarticles/>

²⁵ <http://www.apa.org/pubs/databases/psycbooks/>


²⁶ <http://www.apa.org/pubs/databases/psyc-tests/>

²⁷ <http://www.apa.org/pubs/databases/psyc-therapy/>

the Editor of the Publication Manual of the American Psychological Association²⁸. 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 Science²⁹.

3.21 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 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.

3.22 PsychOpen – The European Open-Access Publishing Platform for Psychology

Erich Weichselgartner (Leibniz Institute for Psychology Information – Trier, DE)

License  Creative Commons BY 3.0 Unported license
© Erich Weichselgartner

The European Psychology Publication Platform PsychOpen was created because extensive research in the European scientific community had clearly shown a demand for open access publishing in psychology. The reasons were manifold. For one, there were only a handful of quality controlled open access journals in psychology in 2011. Secondly, a survey from 493

²⁸ <http://www.apastyle.org/manual/>

²⁹ <https://cos.io/top/>

participants from 24 countries had revealed six main concerns with traditional publishing in psychology: (1) Language, (2) review process, (3) manuscript handling, (4) impact (visibility), (5) permission barriers (accessibility) and (6) price barriers (cost). These issues are the concerns of non-native English speaking Europeans as they experienced in their home countries. PsychOpen was founded in 2013 on the conclusion that an open-access infrastructure would boost scientific and professional communication in European psychology, especially when Europe's language diversity and the lack of resources at the national level (e.g. in Eastern Europe) are taken into account. For the latter reason, to remove hurdles for developing countries and Eastern Europe, but also for strict separation of economic interests and quality control, PsychOpen is Gold Open Access without any author fees.

In order to accomplish its goals efficiently on a small budget, PsychOpen uses a mix of commercial and open source publishing software like PKP's Open Journal System and Inera's eXtyle. Two years after its start, PsychOpen publishes seven journals: Publication languages are English (> 80%), Bulgarian (Non-Roman Script), Portuguese, Spanish and German. The scope is mostly research, one journal is devoted to professional topics. The publication type is traditional research articles. The average publishing time is four months. The publication schedule is continuous in one instance and discrete for the other six journals. Submissions per year and journal range from 25–150; rejection rates are 20%–65%.

All content is published according to the Creative Commons license CC-BY. Two third of PsychOpen authors have a European affiliation; the remaining one third come from North America, East Asia, South America, Africa and South Pacific (in this order). Usage is up by 44% from 2014 to 2015 with approx. 50.000 article downloads in mid-2015. Amongst the challenges for PsychOpen that need further work is multilingualism, the interlinking of scholarly content (e.g., research articles with the corresponding research data), the integration of social media and semantic publishing. A new tool for the semantic enhancement for the Open Journal System facilitates the generation of RDF. Resulting self-describing documents for scientific literature in psychology will allow discovering connections amongst papers and concept-based queries. The lack of ontologies and of NLP tools in psychology, but also the poor data infrastructure are hurdles that need to be overcome.

Participants

- Xavier Aimé
LIMICS INSERM U 1142 –
Paris, FR
- Dietrich Albert
TU Graz, AT
- George Alter
University of Michigan –
Ann Arbor, US
- Björn Brembs
Universität Regensburg, DE
- Mike Conlon
University of Florida, US
- Oscar Corcho
Technical Univ. of Madrid, ES
- Susann Fiedler
MPG – Bonn, DE
- Alexander Garcia Castro
Technical Univ. of Madrid, ES
- Paul Groth
Elsevier Labs – Amsterdam, NL
- William Gunn
Mendeley Ltd. – London, GB
- Janna Hastings
European Bioinformatics
Institute – Cambridge, GB
- Caroline Jay
University of Manchester, GB
- Iris-Tatjana Kolassa
Universität Ulm, DE
- Silvia Koller
Federal University of Rio Grande
do Sul, BR
- Christoph Lange
Universität Bonn, DE
- Maryann Martone
UC – San Diego, US
- Russell Poldrack
Stanford University, US
- Alec Smecher
Simon Fraser University –
Burnaby, CA
- Daniel Staemmler
Elsevier Publishing – Berlin, DE
- Robert Stevens
University of Manchester, GB
- Gary VandenBos
American Psychological
Association, US
- Hal Warren
Vedatek Knowledge Systems, US
- Erich Weichselgartner
ZPID – Trier, DE

