DAGSTUHL
REPORTS

Volume 8, Issue 11, November 2018

**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*
April, 2019

*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`.



Digital Object Identifier: 10.4230/DagRep.8.11.i

*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.



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

Report from Dagstuhl Seminar 18451

# Genomics, Pattern Avoidance, and Statistical Mechanics

**Edited by**

**Michael Albert[1], David Bevan[2], Miklós Bóna[3], and István Miklós[4]**

1 **University of Otago, NZ,** malbert@cs.otago.ac.nz
2 **University of Strathclyde – Glasgow, GB,** david.bevan@strath.ac.uk
3 **University of Florida – Gainesville, US,** bona@ufl.edu
4 **Alfréd Rényi Institute of Mathematics – Budapest, HU,** miklos.istvan.74@gmail.com

## Abstract

We summarize key features of the workshop, such as the three main research areas in which the participants are active, the number and types of talks, and the geographic diversity of the attendees. We also provide a sampling of the collaborations started at the workshop, and explain why we believe that the workshop was successful, and why we believe it should take place again in the future.

**Seminar** November 4–9, 2018 – http://www.dagstuhl.de/18451
**2012 ACM Subject Classification** Mathematics of computing → Approximation algorithms, Applied computing → Bioinformatics, Theory of computation → Data structures design and analysis, Applied computing → Systems biology
**Keywords and phrases** Genome rearrangements, Matrix, Pattern, Permutation, Statistical Mechanics
**Digital Object Identifier** 10.4230/DagRep.8.11.1

## 1 Executive Summary

*Miklós Bóna (University of Florida – Gainesville, US)*



This report documents the program and the outcomes of Dagstuhl Seminar 18451 "Genomics, Pattern Avoidance, and Statistical Mechanics".

The workshop took place from November 4, 2018 to November 9, 2018. It had 40 participants, who were researchers in theoretical computer science, combinatorics, statistical mechanics and molecular biology. It was a geographically diverse group, with participants coming from the US, Canada, Brazil, Germany, Iceland, the United Kingdom, Sweden, France, Switzerland, Hungary, Australia, and New Zealand. The workshop featured 21 talks, three of which were hourlong talks, and an open problem session.

Several collaborative projects have been started. For example, Jay Pantone Michael Albert, Robert Brignall, Seth Pettie, and Vince Vatter started exploring the topic of 1324-avoiding permutations with a bounded number of descents, disproving a 2005 conjecture of Elder, Rechnitzer, and Zabrocki related to Davenport-Schinzel sequences. Had the conjecture been affirmed, it would have implied that the generating function for 1324-avoiding permutations is non-D-finite.

At the open problem session, Yann Ponty raised the following question: what is the number of independent sets in restricted families of trees, like caterpillars or complete binary

plane trees? The main motivation for this question relates to a deep connection between such independent sets and RNA designs. This question led to a new collaborative effort by Mathilde Bouvel, Robert Brignall, Yann Ponty and Andrew Elvey Price.

Sergi Elizalde and Miklós Bóna have started working on Dyck paths that have a unique maximal peak. That collaboration since extended to the area of probabilistic methods, involving a researcher working in that field, Douglas Rizzolo.

Numerous participants expressed their pleasure with the workshop and its sequence of talks. The prevailing view was that while the participants came from three different fields, they were all open to the other two fields, and therefore, they all learned about results that they would not have learned otherwise. Therefore, we have all the reasons to believe that the workshop was a success, and we would like to repeat it some time in the future.

## 2 Table of Contents

# 3 Overview of Talks

## 3.1 The curious behaviour of the total displacement

*David Bevan (University of Strathclyde – Glasgow, GB)*



The *total displacement* of a permutation $\sigma = \sigma_1 \dots \sigma_n$ is $\mathsf{td}(\sigma) = \sum_i |\sigma_i - i|$. The ratio of the total displacement to the number of inversions, $R(\sigma) = \mathsf{td}(\sigma)/\mathsf{inv}(\sigma)$, is known to lie in the half-open interval $(1, 2]$ (unless $\sigma$ is an increasing permutation, with no inversions).

Let $\boldsymbol{\pi}_{n,m}$ denote a permutation chosen uniformly at random from the set of all $n$-permutations with exactly $m$ inversions. In this talk, we consider the behaviour of the expected asymptotic displacement ratio $R[m] = \lim_{n\to\infty} \mathbb{E}\big[R(\boldsymbol{\pi}_{n,m})\big]$, as $m = m(n)$ increases from 1 to $\binom{n}{2}$.

As long as $m = o(n)$, $R[m]$ takes the constant value of 2. Then, when $m \sim \alpha n$, $R[m]$ decreases as $\alpha$ increases, from 2 down to $2\log 2 \approx 1.3863$. However, once $m$ becomes superlinear in $n$, $R[m]$ stalls again, at $2\log 2$, until $m = \Theta(n^2)$. Finally, when $m \sim \rho\binom{n}{2}$, $R[m]$ decreases from $2\log 2$, taking the value $\frac{4}{3}$ when $\rho = \frac{1}{2}$ and finally reaching 1 when $\rho = 1$.

We investigate how this curious behaviour can be explained in terms of the different effects that local and global constraints have on $\boldsymbol{\pi}_{n,m}$.

## 3.2 The solution space of canonical DCJ genome sorting

*Marília Braga (Universität Bielefeld, DE)*


Main reference Marília D. V. Braga, Jens Stoye: "The Solution Space of Sorting by DCJ", Journal of Computational Biology, Vol. 17(9), pp. 1145–1165, 2010.
URL https://doi.org/10.1089/cmb.2010.0109

In genome rearrangements, the double cut and join (DCJ) operation, introduced by Yancopoulos et al. in 2005, allows one to represent most rearrangement events that could happen in multichromosomal genomes, such as inversions, translocations, fusions, and fissions. No restriction on the genome structure considering linear and circular chromosomes is imposed. An advantage of this general model is that it leads to considerable algorithmic simplifications compared to other genome rearrangement models. Several studies about the DCJ operation have been published, and in particular, an algorithm was proposed to find an optimal DCJ sequence for sorting one genome into another one. Here we analyze the solution space of this problem and give an easy-to-compute formula that corresponds to the exact number of optimal DCJ sorting sequences for a particular subset of instances of the problem. We also give an algorithm to count the number of optimal sorting sequences for any instance of the problem. An additional interesting result is the demonstration that, by properly replacing any pair of consecutive operations in any optimal sorting sequence, one always obtains another optimal sorting sequence. As a consequence, any optimal sorting sequence can be obtained from one other by applying such replacements successively.

This work was published in 2010 by Braga and Stoye in the Journal of Computational Biology (DOI: 10.1089/cmb.2010.0109).

## 3.3 Permutations and Permutation Graphs

*Robert Brignall (The Open University – Milton Keynes, GB)*


**Main reference** Nicholas Korpelainen, Vadim V. Lozin, Igor Razgon: "Boundary Properties of Well-Quasi-Ordered Sets of Graphs", Order, Vol. 30(3), pp. 723–735, 2013.
**URL** https://doi.org/10.1007/s11083-012-9272-2

The inversion graph of a permutation provides a convenient tool for translating results and theory between the study of permutations and the study of graphs. In particular, the pattern containment ordering corresponds, in a fairly direct way, to the induced subgraph ordering.

In this talk, I will present two distinct-but-conjecturally-connected topics where permutations have helped: (1) in the study of the graph parameter clique-width, permutation classes provide us with a rich source of hereditary graph properties which are minimal with unbounded clique-width; (2) in the study of well-quasi-ordering for graphs, we exhibit a counterexample to a conjecture made by Korpelainen, Lozin and Razgon, which was found using intuition obtained from the study of well-quasi-ordering in permutations.

## 3.4 On the Median and Small Parsimony problems in some genome rearrangement Models

*Cedric Chauve (Simon Fraser University – Burnaby, CA)*



The main goal of genome rearrangement problems is to compute evolutionary scenarios that can explain the order of genes observed in extant genomes. This naturally leads to questions about the order of genes in ancestral genomes, often of extinct species. If a species phylogeny is given, this problem is known as the Small Parsimony Problem, and in its simplest form, where a single ancestral genome is considered, the Median Problem. In this talk, I will first review several algorithmic results on the Median and Small Parsimony Problems, from initial intractability results to surprising tractability results, and then present some more recent results on the same problems in the context where duplicated genes are considered.

## 3.5 Brain Modularity Mediates the Relation between Task Complexity and Performance

*Michael W. Deem (Rice University – Houston, US)*



Recent work in cognitive neuroscience has focused on analyzing the brain as a network, rather than as a collection of independent regions. Prior studies taking this approach have found that individual differences in the degree of modularity of the brain network relate to performance on cognitive tasks. However, inconsistent results concerning the direction of this relationship have been obtained, with some tasks showing better performance as modularity increases and other tasks showing worse performance. Our recent theoretical model suggests

that these inconsistencies may be explained on the grounds that high-modularity networks favor performance on simple tasks whereas low-modularity networks favor performance on more complex tasks. I will review experiments being carried out by collaborators showing a negative correlation between individuals' modularity and their performance on a composite measure combining scores from the complex tasks and a positive correlation with performance on a composite measure combining scores from the simple tasks. I will further present theory showing that a dynamic measure of brain connectivity termed flexibility is predicted to correlate in the opposite way with performance. I will review experiments confirming these predictions and also showing that flexibility plays a greater role in predicting performance on complex tasks requiring cognitive control and executive functioning. The theory and results presented here provide a framework for linking measures of whole-brain organization from network neuroscience to cognitive processing.

### References


1 A. I. Ramos-Nuez, S. Fischer-Baum, R. Martin, Q.-H. Yue, F.-D. Ye, and M. W. Deem, "Static and Dynamic Measures of Human Brain Connectivity Predict Complementary Aspects of Human Cognitive Performance," *Front. Hum. Neurosci.* (2017) doi: 10.3389/fnhum.2017.00420.

2 Q.-H. Yue, R. Martin, S. Fischer-Baum, A. I. Ramos-Nuez, F.-D. Ye, and M. W. Deem, "Brain Modularity Mediates the Relation of Cognitive Performance to Task Complexity," *J. Cog. Neurosci.* **29** (2017) 1532–1546.

3 J.-M. Park, M. Chen, D. Wang, and M. W. Deem, "Modularity Enhances the Rate of Evolution in a Rugged Fitness Landscape," *Phys. Biol.* **12** (2015) 025001.

4 J.-M. Park, L. R. Niestemski, and M. W. Deem, "Quasispecies Theory for Evolution of Modularity," *Phys. Rev. E* **91** (2015) 012714.


## 3.6 A Markov-chain model of chromosomal instability

*Sergi Elizalde (Dartmouth College – Hanover, US)*


**Joint work of** Sergi Elizalde, Sam Bakhoum, Ashley Laughney

Genomic instability allows cancer cells to rapidly vary the number of copies of each chromosome (karyotype) through chromosome missegregation events during mitosis, enabling genetic heterogeneity that leads to tumor metastasis and drug resistance. We construct a Markov chain that describes the evolution of the karyotypes of cancer cells. The Markov chain is based on a stochastic model of chromosome missegregation which incorporates the observed fact that individual chromosomes contain proliferative and anti-proliferative genes, leading to cells with varying fitness levels and allowing for Darwinian selection to occur. We analyze the Markov chain mathematically, and we use it to predict the long-term distribution of karyotypes of cancer cells. We then adapt it to study the behavior of tumors under targeted therapy and to model drug resistance.

## 3.7 Sampling bipartite degree sequence realizations – the Markov chain approach

*Péter L. Erdös (Alfréd Rényi Institute of Mathematics – Budapest, HU)*


**Joint work of** Péter L. Erdös, Miklós, István
**Main reference** Péter L. Erdös, Tamás Róbert Mezei, István Miklós, Dániel Soltész: "Efficiently sampling the realizations of bounded, irregular degree sequences of bipartite and directed graphs". PLoS ONE 13(8): e0201995, 2018.
**URL** https://doi.org/10.1371/journal.pone.0201995

How to analyze real life networks? There are myriads of them and usually experiments cannot be performed directly on them. Instead, scientists define models, fix parameters and imagine the dynamics of evolution.

Then, they build synthetic networks on this basis (one, several, all) and they want to sample them. However, there are far too many such networks. Therefore, typically, some probabilistic method is used for sampling.

We will survey one such approach, the Markov Chain Monte Carlo method, to sample realizations of given degree sequences. Some new results will be discussed. The majority of the talk is published in [1].

### References

1 P. L. Erdős, T. R. Mezei, I. Miklós, D. Soltész: Efficiently sampling the realizations of bounded, irregular degree sequences of bipartite and directed graphs, *PLOS One 2018* (2018), # e0201995, 1–19.

## 3.8 Sorting Permutations in the GR+IR model

*Guillaume Fertin (University of Nantes, FR)*


**Joint work of** Guillaume Fertin, Géraldine Jean, Eric Tannier
**Main reference** Guillaume Fertin, Géraldine Jean, Eric Tannier: "Genome Rearrangements on Both Gene Order and Intergenic Regions", in Proc. of the Algorithms in Bioinformatics – 16th International Workshop, WABI 2016, Aarhus, Denmark, August 22-24, 2016. Proceedings, Lecture Notes in Computer Science, Vol. 9838, pp. 162–173, Springer, 2016.
**URL** https://doi.org/10.1007/978-3-319-43681-4_13

A genome can be, in its simplest form, modeled as a permutation $\pi$ of length $n$, where each $\pi_i$, $1 \leq i \leq n$, represents a gene. A genome rearrangement (or GR) is a large scale evolutionary event that modifies a genome. For instance, a *reversal* consists in taking a contiguous subsequence from $\pi$, reversing it, and reincorporating it at the same location:

$$\pi = 3\ \underline{5\ 1\ 2}\ 7\ 4\ 6 \rightarrow \pi' = 3\ \underline{2\ 1\ 5}\ 7\ 4\ 6$$

Sorting by rearrangements then consists, given a permutation $\pi$ and a set $\mathcal{S}$ of allowed GRs, in determining a shortest sequence of GRs from $\mathcal{S}$ that transforms $\pi$ in the identity permutation $Id_n$. The algorithmic study of sorting by rearrangements has led to an abundant literature in the last 20 years or so, and given rise to many fascinating results.

It is however possible to enrich the model as follows: since consecutive genes in a genome are actually separated by an *intergenic region* (or IR) – i.e. by a certain number of DNA bases –, we can model a genome by a pair consisting of (a) a permutation $\pi$, together with (b)

an ordered multiset $S = \{r_1, r_2 \cdots r_{n-1}\}$ of positive integers, where each $r_j$, $1 \leq j \leq n-1$, represents the size of the IR between genes $\pi_j$ and $\pi_{j+1}$. A GR acts between genes, thus inside IRs – in the above example, the shown reversal cuts $\pi$ between genes 3 and 5, and between genes 2 and 7. Hence, any GR can simultaneously modify the sizes of the affected IRs *and* the order of the genes.

In this setting, which we call the GR+IR model, the sorting problem becomes the following: given a pair $(\pi, S)$ representing a genome together with its IRs, find a shortest sequence of GRs that leads to the pair $(Id_n, S')$, where $S'$ encodes the IRs of the target permutation.

In this talk, I will introduce the GR+IR model in more details, and give some algorithmic results related to the corresponding sorting problem, with a specific focus on the following two types of GR: DCJ (Double Cut and Join) and reversals.

## 3.9 The combinatorics of RNA branching

*Christine E. Heitsch (Georgia Institute of Technology – Atlanta, US)*



Understanding the folding of RNA sequences into three-dimensional structures is one of the fundamental challenges in molecular biology. For example, the branching of an RNA secondary structure is an important molecular characteristic yet difficult to predict correctly. However, results from enumerative, probabilistic, and geometric combinatorics can characterize different types of branching landscapes, yielding insights into RNA structure formation.

## 3.10 Modelling dense polymers by self-avoiding walks

*E. J. Janse van Rensburg (York University — Toronto, CA)*



A dense polymer (for example in a polymer melt, or polymers in confined spaces in living cells) can be modelled by a lattice self-avoiding walk in a confined space. For example, in the square lattice a self-avoiding walk can be confined to a square. If the walk is very short compared to size of the square, then it is in an expanded phase, but when it is long, then it will start to fill the area of the square and is a compressed walk. In this talk I give a summary about modelling the free energy of a compressed walk by using Flory-Huggins theory (a mean field phenomenological theory of the free energy of dense polymer systems). We estimate numerically the Flory Interaction Parameter for square lattice self-avoiding walks, and also give an extrapolated estimate of the connective constant of Hamiltonian walks of a square. I will also produce tentative results on compressed and knotted lattice polygons in 3 dimensions.

## 3.11 Pattern avoidance: algorithmic connections

*László Kozma (FU Berlin, DE)*



Permutation-patterns and pattern-avoiding permutations have long been the subject of algorithmic study. Classical problems include pattern matching, enumeration of pattern-occurrences and of pattern-avoiding sequences, and others. More recently, pattern-avoidance was also studied in the property-testing framework.

Perhaps less-studied – apart from special cases – is the question whether the avoidance of patterns makes classical algorithmic problems easier. (Analogous questions in graphs are well-studied.) In my talk I discuss sorting and searching in basic data structures, when the input is pattern-avoiding.

## 3.12 Combinatorics of the ASEP on a ring and Macdonald polynomials

*Olya Mandelshtam (Brown University – Providence, US)*


**Joint work of** Olya Mandelshtam, Sylvie Corteel, Lauren Williams
**Main reference** Sylvie Corteel, Olya Mandelshtam, Lauren Williams: "Combinatorics of the two-species ASEP and Koornwinder moments". Advances in Mathematics, Vol. 321, pp.160–204, 2017
**URL** https://doi.org/10.1016/j.aim.2017.09.034

The multispecies asymmetric simple exclusion process (ASEP) is a model of hopping particles of $M$ different types hopping on a one-dimensional lattice of $N$ sites. In this talk, we consider the ASEP on a ring with the following dynamics: particles at adjacent sites can swap places with either rate 1 or $t$ depending on their relative types. Recently, James Martin gave a combinatorial formula for the stationary probabilities of the ASEP with generalized *multiline queues*. We will begin by describing the combinatorial methods we use to study the ASEP on a ring.

Furthermore, it turns out that by introducing additional statistics on the multiline queues, we get a new formula for both symmetric Macdonald polynomials $P_\lambda$ and nonsymmetric Macdonald polynomials $E_\lambda$, where $\lambda$ is a partition. For the second part of the talk, we will discuss the recent results and remarkable connection with Macdonald polynomials.

## 3.13 Computational complexity of counting and sampling genome rearrangement scenarios

*István Miklós (Alfréd Rényi Institute of Mathematics – Budapest, HU)*


**Joint work of** István Miklós, Heather C. Smith, Eric Tannier
**Main reference** István Miklós, Heather Smith: "Sampling and counting genome rearrangement scenarios", BMC Bioinformatics, Vol. 16(Suppl 14), pp. S6, 2015.
**URL** https://doi.org/10.1186/1471-2105-16-S14-S6

Most of the counting problems fall into one of the following 3 categories:
1. In FP, that is, exactly solvable in polynomial time

2. In the intersection of #P-complete and FPRAS, that is, exact polynomial solution does not exists (assuming that P $\neq$ NP) but efficient random approximation exists
3. In #P-complete and outside of FPRAS, that is, cannot be well approximated even in a stochastic manner (assuming that RP $\neq$ NP).

The sampling counterparts usually follow the counting complexity and thus there is a perfect uniform sampler, an approximate uniform sampler or any sampler that runs in polynomial time is far from the uniform distribution.

There are several genome rearrangement models (sorting by reversals, SCJ, DCJ, etc.) and several tasks (counting the most parsimonious scenarios between two genomes, computing the number of median genomes, etc.), and for each combination of models and tasks, we are interested in the computational complexity of the so-emerging computational problem. The talk will focus on the recent results and open problems. Connections to enumerative combinatorics, statistical physics and network analysis will also be discussed.

## 3.14 On some generalizations to trees of problems for permutations

*Alois Panholzer (TU Wien, AT)*


**Main reference** Marie-Louise Lackner, Alois Panholzer: "Parking functions for mappings", J. Comb. Theory, Ser. A, Vol. 142, pp. 1–28, 2016.
**URL** https://doi.org/10.1016/j.jcta.2016.03.001

Various enumeration problems and statistics for permutations have been generalized to other combinatorial structures. In this talk we focus on some of such generalizations to labelled tree structures. In particular, some old and new results for random sequential adsorption, records, and local label-patterns in trees are discussed.

## 3.15 Sorting Permutations with C-Machines

*Jay Pantone (Marquette University, US), Michael Albert (University of Otago, NZ), Cheyne Homberger, Nathaniel Shar, and Vincent Vatter (University of Florida – Gainesville, US)*


**Joint work of** Michael H. Albert, Cheyne Homberger, Jay Pantone, Nathaniel Shar, Vincent Vatter
**Main reference** Michael H. Albert, Cheyne Homberger, Jay Pantone, Nathaniel Shar, Vincent Vatter: "Generating permutations with restricted containers", J. Comb. Theory, Ser. A, Vol. 157, pp. 205–232, 2018.
**URL** https://doi.org/10.1016/j.jcta.2018.02.006

A C-machine is a type of sorting device that naturally generalizes stacks and queues. A C-machine is a container that is allowed to hold permutations from the permutation class C into which entries can be pushed and out of which entries may be popped. With this notation, a traditional stack is the Av(12)-machine. This structural description allows us to find many terms in the counting sequences of several permutation classes of interest, but despite these numerous initial terms we are unable to find the exact or asymptotic behavior of their generating functions. I'll discuss what we do know, what we don't know, and what experimental methods tell us we might one day know.

## 3.16 Amortized Analysis of Data Structures via Forbidden 0-1 Matrices

*Seth Pettie (University of Michigan – Ann Arbor, US)*


**Main reference** Seth Pettie: "On Nonlinear Forbidden 0-1 Matrices: A Refutation of a Füredi-Hajnal Conjecture", in Proc. of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010, pp. 875–885, SIAM, 2010.
**URL** https://doi.org/10.1137/1.9781611973075.71

The amortized performance of a data structure is usually proved by designing and analyzing a "potential function", which is an accounting mechanism for letting faster-than-average operations pay for slower-than-average operations. In this talk I will survey an alternative method for analyzing amortized data structures that models executions by 0-1 matrices and bounds their weight using theorems on the density of such matrices avoiding 0-1 patterns.

## 3.17 Lattice Path Counting: where Enumerative Combinatorics and Statistical Mechanics meet

*Thomas Prellberg (Queen Mary University of London, GB)*



A topic common to the two disciplines in the title of this talk is the wish to count truly large ensembles of structures. This talk will examine different ways of how the problem of lattice path counting is approached using methods from both of these areas. While enumerative combinatorics strives to ideally provide exact counting numbers, statistical mechanics rather deals with the thermodynamic limit of large system sizes, where concepts like entropy and energy are related to asymptotic growth. I will endeavour to close the gap between exact counting formulas from enumerative combinatorics and the approximate counting underlying the statistical mechanical approach, and to clearly define jargon particular to either discipline.

## 3.18 Permutations in labellings of trees

*Fiona Skerman (Uppsala University, SE)*


**Joint work of** Fiona Skerman, Michael Albert, Cecilia Holmgren, Tony Johansson

We investigate the number of permutations that our in random node labellings of trees. This is a generalisation of the number of sub-permutations occurring in a random permutation. It also generalises some recent results on the number of inversions in randomly labelled trees by Cai, Holmgren, Janson, Johansson and Skerman. We consider complete binary trees as well as random split trees a large lass of random trees of logarithmic height introduced by Devroye. Split trees consist of nodes (bags) which an contain balls and are generated by a random trickle down process of balls through the nodes.

In the case of the complete binary trees that asymptotically the cumulants of the number of occurrences of a fixed permutation in the random node labelling have explicit formulas. For a random split tree with high probability we show the cumulants of the number of

occurrences are asymptotically an explicit parameter of the split tree. I will describe some results on the number of embeddings of digraphs into split trees, used in the proof of the second result, which may be of independent interest.

This is joint work with Michael Albert, Cecilia Holmgren, Tony Johansson.

## 3.19 Enumerating $1 \times n$ generalised permutation grid classes

*Jakub Sliacan (University of Umeå, SE) and Robert Brignall (The Open University – Milton Keynes, GB)*



This talk is concerned with the study of $1 \times n$ almost-monotone permutation grid classes. In particular, we show that $1 \times n$ grid classes with $n - -1$ monotone cells and one context-free cell admit algebraic generating functions. Our approach is algorithmic and, in principle, allows us to enumerate any such class in particular (assuming sufficient computational resources). We give examples to illustrate the method on familiar objects.

Our methods, which leverage the inductive/recursive structure of these $1 \times n$ classes, will be the focus of the talk. We rely on combinatorial specifications of context-free classes and define operators on them which do the job of "appending a monotone class to the right of a context-free cell". With appropriate pre- and post-processing, this constitutes the method in its entirety.

## 3.20 Complexity of the Single Cut-or-Join model and Partition Functions

*Heather Smith (Davidson College, US) and István Miklós (Alfréd Rényi Institute of Mathematics – Budapest, HU)*


**Main reference** István Miklós, Heather Smith: "The computational complexity of calculating partition functions of optimal medians with Hamming distance", Advances in Applied Mathematics, Vol. 102, pp.18–82, 2019
**URL** https://doi.org/10.1016/j.aam.2018.09.002

We survey computational complexity results for the Single Cut-or-Join model for genome rearrangement, a common mode of molecular evolution. Our main result, enumerating the most parsimonious median scenarios is #P-complete, follows from a more general result for partition functions. In particular, calculating the partition function of optimal medians of binary strings with Hamming distance is #P-complete for several weight functions. This is joint work with István Miklós.

## 3.21 Two topics on permutation patterns

*Vincent Vatter (University of Florida – Gainesville, US), Michael Engen, and Jay Pantone (Marquette University, US)*


**Main reference** Michael Engen, Vincent Vatter: "Containing all permutations", CoRR, Vol. abs/1810.08252, 2018.
**URL** https://arxiv.org/abs/1810.08252
**Main reference** Jay Pantone, Vincent Vatter: "Growth rates of permutation classes: categorization up to the uncountability threshold", CoRR, Vol. abs/1605.04289, 2016
**URL** https://arxiv.org/abs/1605.04289
**Main reference** Vincent Vatter: "Growth rates of permutation classes: from countable to uncountable", CoRR, Vol. abs/1605.04297, 2016
**URL** https://arxiv.org/abs/1605.04297

In this talk I will discuss two (admittedly unconnected) aspects of permutation patterns. First, I will discuss the determination of the set of all growth rates of permutation classes. Recently, in joint work with Pantone, we have increased the classification of these growth rates up to approximately 2.30, which is the point at which there begin to be uncountably many such growth rates. Given that Bevan has previously shown that all real numbers greater than approximately 2.36 are growth rates of permutation classes, the gap within which these growth rates remain unclassified is only 0.06 wide.

Secondly, I will discuss various versions of the problem of "containing all permutations". In one version of this problem, Miller had shown in 2009 that there is a permutation of length $(n^2+n)/2$ which contains all permutations of length $n$ as subsequences. I will discuss how Engen and I have recently lowered this bound to $\lceil (n^2+1)/2 \rceil$. I then discuss another version of this problem, which has attracted media attention from such outlets as *The Verge*, *Quanta Magazine*, and *Wired*. In this version of the problem, we must contain all permutations of length $n$, but contiguously, not as factors, and the object that is to contain them must be a word over the same alphabet $\{1, 2, \ldots, n\}$. There was a long-standing conjecture that the answer in this case was $n! + (n-1)! + \cdots + 3! + 2! + 1!$. This was disproved by Houston in 2014, who constructed such a "superpermutation" in the $n = 6$ case which had length only 872 (one less than the conjectured length). Then, last month, the science fiction author Greg Egan unveiled a construction of such a superpermutation of length only $n! + (n-1)! + (n-2)! + (n-3)! + n - 3$. The best lower bound to-date (which was posted anonymously on the website *4Chan* in 2011 but not read carefully until the recent new interest in the problem) is $n! + (n-1)! + (n-2)! + n - 3$, leaving a gap of only $(n-3)!$.

## 3.22 On Square Permutations

*Stéphane Vialette (University Paris-Est – Marne-la-Vallée, FR)*



Given permutations $\pi$ and $\sigma_1$ and $\sigma_2$, the permutation $\pi$ is said to be a *shuffle* of $\sigma_1$ and $\sigma_2$, in symbols $\pi \in \sigma_1 \,\Delta\, \sigma_2$, if $\pi$ (viewed as a string) can be formed by interleaving the letters of two strings $p_1$ and $p_2$ that are order-isomorphic to $\sigma_1$ and $\sigma_2$, respectively. In case $\sigma_1 = \sigma_2$, the permutation $\pi$ is said to be a *square* and $\sigma_1 = \sigma_2$ is a *square root* of $\pi$. For example, $\pi = 24317856$ is a square as it is a shuffle of the patterns 2175 and 4386 that are both order-isomorphic to $\sigma = 2143$ as shown in $\pi = {}^{2}{}_{43}{}^{17}{}_{8}{}^{5}{}_{6}$. However, $\sigma$ is not the

unique square root of $\pi$ since $\pi$ is also a shuffle of patterns 2156 and 4378 that are both order-isomorphic to 2134 as shown in $\pi = 2_{43}1_{78}56$.

We shall begin by presenting recent results devoted to recognizing square permutations and related concepts with a strong emphasis on constrained oriented matchings in graphs. Then we shall discuss research directions to address square permutation challenges in both combinatorics and algorithmic fields.

# 4 Working groups

## 4.1 Enumerative aspects of multiple RNA design

*Mathilde Bouvel (Universität Zürich, CH), Robert Brignall (The Open University – Milton Keynes, GB), Andrew Elvey Price (The University of Melbourne, AU), and Yann Ponty (Ecole Polytechnique – Palaiseau, FR)*



In this note, we compute the number of independent sets on certain graphs: unions of paths and cycles, caterpillars (a.k.a. combs), and complete binary trees. This question was raised during the open problem session of the Dagstuhl Seminar 18451 "Genomics, Pattern Avoidance, and Statistical Mechanics", in relation with applications to RNA design. Indeed, RNA sequences compatible with a set of RNA secondary structures are in correspondence with independent sets on the dependency graph of this set of structures.

### Context

At the open problem session of the Dagstuhl Seminar 18451 "Genomics, Pattern Avoidance, and Statistical Mechanics", Yann Ponty raised the following question: what is the number of independent sets in restricted families of trees, like caterpillars or complete binary plane trees? The main motivation for this question relates to a deep connection between such independent sets and RNA designs, that we elaborate below.

### Definitions and problem statement

An *RNA secondary structure* $S$ of length $n$ is a set of base pairs $\{(i, j)\}$ such that $1 \leq i < j \leq n$, such that each position in $[1, n]$ is involved in at most a single base pair. Typical definitions for the secondary structure include additional constraints, for instance to preclude crossing base pairs or to ensure a minimal distance between paired positions, which we omit here for the sake of simplicity.

An RNA sequence $w \in \{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}^n$ is *compatible* with a single secondary structure $S$ of length $n$ when

$$\forall (i, j) \in S : \{w_i, w_j\} \in \mathcal{B}, \text{where } \mathcal{B} := \{\{\mathsf{G}, \mathsf{C}\}, \{\mathsf{A}, \mathsf{U}\}, \{\mathsf{G}, \mathsf{U}\}\}.$$

We say that a sequence is compatible with a set of structures $\mathcal{S}$ (all of the same length) when it is compatible with each structure $S \in \mathcal{S}$.

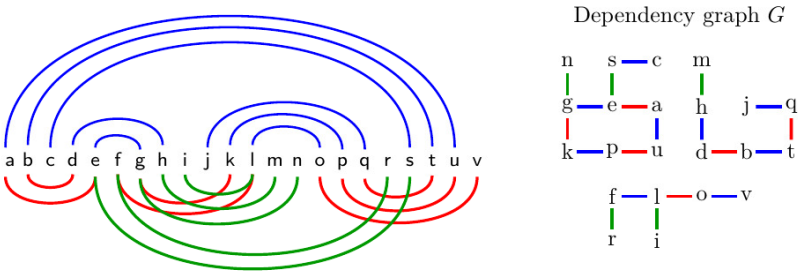



**Figure 1** Three secondary structures and their associated dependency graph.

Given a set $\mathcal{S}$ of structures all of length $n$, the set of RNA sequences compatible with $\mathcal{S}$ depends only on the *(undirected labeled) dependency graph* $G$ of $\mathcal{S}$, which is defined as $G = (V, E)$ with $V = [1, n]$ and $E = \cup_{S \in \mathcal{S}} \{(i, j) \in S\}$. The set of RNA sequences compatible with $\mathcal{S}$ is denoted by $\mathsf{Design}(G)$.

We are interested in finding computable expressions for the number of RNA sequences compatible with a given $\mathcal{S}$, of dependency graph $G$. This number is by definition

$$\#\mathsf{Design}(G) = \left|\{w \in \{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}^n \mid \forall S \in \mathcal{S}, w \text{ compatible with } S\}\right|.$$

Note that, whenever $G$ admits an odd cycle, it is impossible to assign letters to $G$ in a way that fulfills the compatibility requirement. It follows that $\#\mathsf{Design}(G) = 0$ for any non-bipartite graph $G$, thus we restrict our scope to bipartite graphs.

▶ **Lemma 1.** *Denote by $\mathcal{C}_G$ the set of connected components of a bipartite graph $G$. One has*

$$\#\mathsf{Design}(G) = 2^{|\mathcal{C}_G|} \times \prod_{c \in \mathcal{C}_G} \#\mathsf{IndSets}(c),$$

*where $\#\mathsf{IndSets}(c)$ is the number of independent sets of a (connected) graph $c$.*

**Proof.** Assume first that $G$ is connected. Given a compatible labeling of the vertices of $G$ by letters in $\{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}$, the set of vertices labeled by $A$ or $C$ forms an independent set, since $\{\mathsf{A}, \mathsf{C}\} \notin \mathcal{B}$. Conversely, given an independent set $I$ of $G$, we build an RNA sequence compatible with $G$ by assigning letters in $\{\mathsf{A}, \mathsf{U}, \mathsf{C}, \mathsf{G}\}$ to the vertices as follows:

- vertices in $I$ are assigned $A$ or $C$;
- vertices not in $I$ are assigned $U$ or $G$;
- choose the label of the vertex 1 (among two possibilities as above, depending on whether $1 \in I$ or not).

Because $G$ is connected, once the label of the vertex 1 has been chosen, then all other labels are determined by the fact that all edges need to be labeled in accordance with $\mathcal{B} = \{\{\mathsf{G}, \mathsf{C}\}, \{\mathsf{A}, \mathsf{U}\}, \{\mathsf{G}, \mathsf{U}\}\}$. This results in a two-to-one correspondence between $\mathsf{Design}(G)$ and the set of independent sets of $G$.

This immediately extends to graphs $G$ with several connected components: in this case, we just need to choose (among two possibilities) the label of a vertex in each connected component. ◀

Note that computing $\#\mathsf{IndSets}(G)$ is a well-studied #P-hard problem, even for graphs of maximum degree 3 [1]. Since any bipartite graph $G$ with $n$ vertices can be obtained as the dependency graph of $\Theta(n)$ secondary structures, computing $\#\mathsf{Design}$ is #P-hard in general,

yet solvable in time polynomial time for dependency graphs of bounded tree width [2]. Given the hardness of the general problem, and its practical relevance to applications based on random generation, we consider enumerative properties of simple classes of dependency graphs.

## 5 Explicit enumeration results on constrained dependency graphs

A discussion between the four authors of this note resulted in the following results, giving formulas for $\#\mathsf{IndSets}(G)$ (and hence $\#\mathsf{Design}(G)$) when $G$ is a union of paths and cycles, a caterpillar or a complete binary tree.

### 5.1 Union of paths and cycles

By definition, when $G$ is the dependency graph of a set $\mathcal{S}$ containing two structures, every vertex of $G$ has degree at most two. Thus, every connected component of $G$ is either a path or a cycle.

▶ **Lemma 1.** *The number $p_n$ of independent sets of a path with $n$ vertices satisfies the recurrence*

$$p_n = p_{n-1} + p_{n-2},$$

*with initial conditions $p_0 = 1$, $p_1 = 2$ i.e., $p_n$ is the $(n+2)$-th Fibonacci number ([3, A000045]).*

**Proof.** Let $P$ be a path with $n$ vertices. Consider a vertex $v$ at an extremity of the path. There are $p_{n-1}$ independent sets of $P$ not containing $v$, and $p_{n-2}$ which do contain $v$. ◀

▶ **Lemma 2.** *The number $c_n$ of independent sets of a cycle with $n$ vertices is*

$$c_n = p_{n-1} + p_{n-3}.$$

**Proof.** Let $C$ be a cycle with $n$ vertices, and $v$ be a vertex (for instance, the one with label 1). The number of independent sets of $C$ which do not contain $v$ is $p_{n-1}$, and then number of those containing $v$ is $p_{n-3}$. ◀

Putting Lemmas 1, 1 and 2 together, we obtain the following.

▶ **Proposition 1.** *Let $f_n$ be the $n$-th Fibonacci number, defined by $f_0 = 0$, $f_1 = 1$ and $f_{n+2} = f_{n+1} + f_n$. The number of designs of a dependency graph $G$ associated with a set $\mathcal{S}$ of two structures is given by*

$$\#\mathsf{Design}(G) = \prod_{p \in \mathcal{C}_G \text{ is a path}} 2\, f_{|p|+2} \times \prod_{c \in \mathcal{C}_G \text{ is a cycle}} \left(2\, f_{|c|+1} + 2\, f_{|c|-1}\right).$$

### 5.2 Caterpillars

We define a family of binary trees which we call *caterpillars* (and are sometimes called combs or centipede graphs in the literature) as follows. There is exactly one caterpillar of each

size $n \geq 0$. The caterpillar of size 0 is the tree with just one vertex. For any $n \geq 1$, the caterpillar of size $n$ is the tree whose root has two children, the left child being a leaf, and the right child being the caterpillar of size $n-1$.

Note that these trees are rooted, plane, and unlabeled.

The caterpillars of size 0 to 2 are shown in Figure 2.

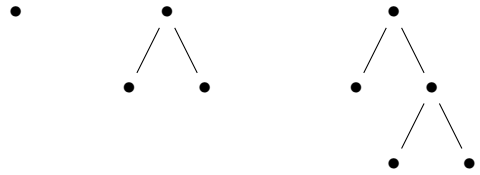


**Figure 2** The caterpillars of size 0 to 2.

▶ **Lemma 1.** *The number $a_n$ of independent sets of the caterpillar of size $n$ satisfies the recurrence*

$$a_n = 2a_{n-1} + 2a_{n-2}$$

*with initial conditions $a_0 = 2$, $a_1 = 5$. Denoting $(a'_n)$ the sequence [3, A052945], we have $a_n = a'_{n+1}$.*

**Proof.** Let $G$ be the caterpillar of size $n$, and let $v$ be the root of $G$. Let also $\ell$ be the left child of $v$, $r$ be its right child, and $u_L$ and $u_R$ be the left and right children of $r$, respectively. An independent set $I$ of $G$ may or not contain $v$.

- If $I$ does not contain $v$, then $\ell$ may or not be in $I$, and $I$ restricted to the subtree rooted at $r$ is a generic independent set of a caterpillar of size one less. So, there are $2a_{n-1}$ independent sets of $G$ which do not contain $v$.
- If $I$ contains $v$, then $\ell \notin I$ and $r \notin I$. But then, similarly to the above case, $u_L$ may or not be in $I$, and $I$ restricted to the subtree rooted at $u_R$ is a generic independent set of a caterpillar of size two less. So, there are $2a_{n-2}$ independent sets of $G$ which contain $v$. ◀

Combining Lemmas 1 and 1 yields the following.

▶ **Proposition 2.** *Let $G$ be a (labeled) dependency graph with $2n+1$ vertices which admits a rooting and a planar embedding that allows to see $G$ as a labeled caterpillar (necessarily of size $n$). The number of designs of $G$ is $2a_n$.*

## 5.3 Complete binary trees

We consider now another family of trees containing one tree of each size: namely, the complete binary trees (where the size is the depth). Again, these trees are rooted, plane, and unlabeled. Examples are shown in Figure 3.

▶ **Lemma 1.** *The number $b_n$ of independent sets of the complete binary tree of size $n$ satisfies the recurrence*

$$b_n = b_{n-1}^2 + b_{n-2}^4$$

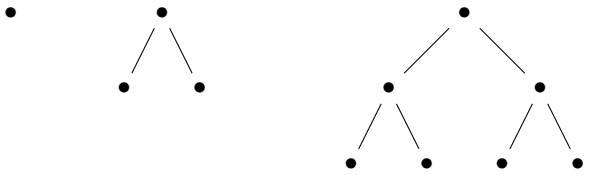


**Figure 3** The complete binary trees of size 0 to 2.

*with initial conditions* $b_0 = 2$, $b_1 = 5$. *Denoting* $(b'_n)$ *the sequence [3, A052945], we have* $b_n = b'_{n+2}$.

**Proof.** The formula follows as in the previous proofs, the term $b^2_{n-1}$ (resp. $b^4_{n-2}$) counting the independent sets which do not (resp. do) contain the root of the binary tree. ◀

From the above and Lemma 1, we have:

▶ **Proposition 3.** *Let* $G$ *be a (labeled) dependency graph with* $2^{n+1} - 1$ *vertices which admits a rooting and a planar embedding that allows to see* $G$ *as a labeled complete binary tree (necessarily of size* $n$*). The number of designs of* $G$ *is* $2b_n$.

## 6 Going further

This note just records the results of a discussion during the seminar. However, it will hopefully serve as a basis for starting a collaboration. The following questions may be of interest.

- Find enumerative information on the sequences $a_n$ and $b_n$, like a closed form or asymptotic behavior. Note that it will just be a routine exercise to obtain this information for $a_n$, but is not immediate for the case of $b_n$.
- Compute $\#\mathsf{Design}(G)$ in other cases, starting with other families of trees. In particular, consider some families of trees which contain *several* trees of a given size, and express $\#\mathsf{Design}(G)$ by a formula involving not only the size but also the value of a parameter to determine.

### References

1 Martin Dyer and Catherine Greenhill. On markov chains for independent sets. *Journal of Algorithms*, 35(1):17 – 49, 2000. http://dx.doi.org/https://doi.org/10.1006/jagm.1999.1071

2 Stefan Hammer, Yann Ponty, Wei Wang, and Sebastian Will. Fixed-Parameter Tractable Sampling for RNA Design with Multiple Target Structures. In *RECOMB 2018 – 22nd Annual International Conference on Research in Computational Molecular Biology*, Paris, France, April 2018. Extended version under review. URL: https://hal.inria.fr/hal-01631277.

3 OEIS Foundation Inc. The encyclopedia of integer sequences, 2011. URL: http://oeis.org.

## Participants

- Michael Albert
University of Otago, NZ
- David Bevan
University of Strathclyde – Glasgow, GB
- Miklós Bóna
University of Florida – Gainesville, US
- Mathilde Bouvel
Universität Zürich, CH
- Marilia Braga
Universität Bielefeld, DE
- Robert Brignall
The Open University – Milton Keynes, GB
- Cedric Chauve
Simon Fraser University – Burnaby, CA
- Anders Claesson
University of Iceland – Reykjavik, IS
- Michael W. Deem
Rice University – Houston, US
- Sergi Elizalde
Dartmouth College – Hanover, US
- Andrew Elvey Price
The University of Melbourne, AU
- Péter L. Erdös
Alfréd Rényi Institute of Mathematics – Budapest, HU
- Guillaume Fertin
University of Nantes, FR
- Yoong Kuan Goh
University of Technology – Sydney, AU
- Torin Greenwood
Rose-Hulman Inst. of Technology – Terre Haute, US
- Sylvie Hamel
Université de Montréal – Québec, CA
- Christine E. Heitsch
Georgia Institute of Technology – Atlanta, US
- E. J. Janse van Rensburg
York University – Toronto, CA
- László Kozma
FU Berlin, DE
- Anthony Labarre
University Paris-Est – Marne-la-Vallée, FR
- Olya Mandelshtam
Brown University – Providence, US
- István Miklós
Alfréd Rényi Institute of Mathematics – Budapest, HU
- Alois Panholzer
TU Wien, AT
- Jay Pantone
Marquette University, US
- Seth Pettie
University of Michigan – Ann Arbor, US
- Yann Ponty
Ecole Polytechnique – Palaiseau, FR
- Svetlana Poznanovik
Clemson University, US
- Thomas Prellberg
Queen Mary University of London, GB
- Pijus Simonaitis
University of Montpellier 2, FR
- Fiona Skerman
Uppsala University, SE
- Jakub Sliacan
University of Umeå, SE
- Heather Smith
Davidson College, US
- Jason Smith
University of Aberdeen, GB
- Rebecca Smith
The College at Brockport, US
- Einar Steingrimsson
University of Strathclyde – Glasgow, GB
- Jens Stoye
Universität Bielefeld, DE
- Jessica Striker
North Dakota State University – Fargo, US
- Krister Swenson
University of Montpellier & CNRS, FR
- Vincent Vatter
University of Florida – Gainesville, US
- Stéphane Vialette
University Paris-Est – Marne-la-Vallée, FR

Report from Dagstuhl Seminar 18461

# Blockchain Security at Scale

**Edited by**

**Rainer Böhme[1], Joseph Bonneau[2], and Ittay Eyal[3]**

1 **Universität Innsbruck, AT,** rainer.boehme@uibk.ac.at
2 **New York University, US,** jbonneau@gmail.com
3 **Technion – Haifa, IL,** ittay@technion.ac.il

## Abstract

38 researchers affiliated with over 25 different institutions in 7 countries met during Dagstuhl Seminar 18461 for discussing open problems regarding "Blockchain Security at Scale." The seminar was split into eight blocks of two presentations each. The mode for each talk was 15 minutes of blackboard-only presentation followed by 30 minutes of discussion. Discussions not fitting into this limit were resumed in smaller break-out groups. This report documents the scheduled talks as well as the improvised sessions for in-depth discussion.

**Seminar** November 11–16, 2018 – http://www.dagstuhl.de/18461
**2012 ACM Subject Classification** Networks → Network security, Computer systems organization → Peer-to-peer architectures
**Keywords and phrases** Blockchain, Consensus, Cryptography, Distributed Systems, Game Theory, Scaling
**Digital Object Identifier** 10.4230/DagRep.8.11.21
**Edited in cooperation with** Assimakis Kattis, Patrik Keller, Itay Tsabary

## 1 Executive Summary

*Rainer Böhme (Universität Innsbruck, AT)*



The security of blockchain-based systems has attracted great interest in the research community following the initial financial success of Bitcoin. Several security notions for blockchain-based systems have been proposed, varying in degree of formality and applicability to real-world systems. However, a major blind spot remains about the environment surrounding blockchain-based systems. This environment is typically assumed to be static (irresponsive to activities of the blockchain system). This is a sound starting point for security analysis while the stakes involved are small compared to the environment (i. e., the global economic and political system). However, if blockchain-based systems truly offer compelling advantages over legacy systems, they may eventually become the dominant form of organizing certain social choice problems. This "scale change" challenges the assumption that the blockchain-based system remains below the threshold of relevance for the parts of its environment that are vital for its security. One instance where this may already occur is the influence of mining puzzles on hardware design and electricity prices.


Blockchain Security at Scale, *Dagstuhl Reports*, Vol. 8, Issue 11, pp. 21–34
Editors: Rainer Böhme, Joseph Bonneau, Ittay Eyal
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The purpose of the seminar was to bring together researchers with expertise in various subfields of blockchain-based systems to jointly revisit security foundations. The primary goal was to incorporate explicit consideration of reciprocity effects between properties of cryptocurrency protocols and their environment.

The primary intended outcome of this seminar was proposing a new design principle, viewing security as a key scalability property to consider in addition to performance and efficiency. Second, the seminar aimed to converge on standard terminology for security notions that are robust to scale. Third, we applied this new methodology to Bitcoin specifically as a test case, producing a sort-of "break glass in case of rampant runaway growth" security plan.

Specific questions were:

1. **From micro-level to macro-level incentives** Bitcoin's ecosystem remains small relative to large multinational corporations. What happens to incentives when a cryptocurrency reaches a scale similar to large national economies?
2. **Cryptographic agility** How does the ability to upgrade cryptographic algorithms might change in the future as cryptocurrency protocols become widely embedded in hardware and/or codified in the law.
3. **Reciprocity effects on hardware design** How will the hardware industry be affected by the increasing importance of superior hardware for mining, and possibly trusted execution environments (TEE) in the future?
4. **Mining economics at scale** How will mining economics change in the future, in particular, dynamics between miners at large-scale power consumption levels, with mass availability of cheap commodity mining hardware (including TEE-based), and with different incentives, e.g., in a high-valued fee-only revenue model.
5. **Reconsidering non-monetary incentives** Can cryptocurrencies be resilient to disruptive nation-level attacks that are not due to monetary incentives?
6. **Governance at scale** To date, cryptocurrencies largely rely on informal leadership from a small group of influential software developers. Can this be translated into a more democratic model? What does democratic control mean for a cryptocurrency when the *demos* is not clearly defined?

## 2 Table of Contents

# 3 Overview of Talks

## 3.1 Using Differential Privacy to Analyze Cryptocurrencies Anonymity

*Foteini Baldimtsi (George Mason University – Fairfax, US), transcription from abstract book*



We investigate whether techniques inspired from the area of differential privacy can be used to construct anonymous cryptocurrencies offering an interesting set of trade-offs between the level of offered privacy, efficiency and underlying assumptions.

Our motivation rises from the fact that Monero, one of the most popular private cryptocurrencies has been recently analyzed to find that approximately 60% of transactions provide no or very limited privacy due to a very small anonymity set. We propose a protocol inspired by Monero (utilizing ring signatures) and formally analyze it while preserving differential privacy for users. Specifically we would like to claim that two neighboring transaction graphs are nearly equal to give rise to the same chain. In order to keep the size of each individual ring signature small while providing a large number of potential options for the real transaction we have users submit several ring signatures in a sequence of rounds that "pipe-in" an ever-increasing number of possible mix-ins.

## 3.2 STARKs for Blockchain Scalability

*Eli Ben-Sasson (Technion – Haifa, IL)*



An interactive proof system is defined to be a STARK if it satisfies the following conditions:

Scalability: for statements referring to computations of nondeterministic time T, proving time scales quasi-linearly in T and verification time scales poly-logarithmically in T.

Transparency: all verifier messages are public random coins ARgument of Knowledge: There exists a polynomial time extractor that, interacting with a valid prover, reconstructs a non-deterministic witness for the statement.

STARKs are unique in their scalability capabilities, when compared to the zk-SNARKs deployed in Zcash, the recursive SNARKs suggested for Coda and the BulletProofs system used by Monero. The talk described the essential properties of STARKs and compared them to SNARKs, recursive SNARKs and BulletProofs.

## 3.3 Proof of Work and Resource Hardness

*Alex Biryukov (University of Luxembourg, LU), transcription from abstract book*



We have defined the properties a resource hard proof of work should have. We have defined classed of $R$-hardness depending on prover-verifier capabilities (hard, easy with secret, publicly easy) and with regard to the specific resource $R$: Time (sequential or total computation),

memory, code size. In this setting PoWs in the Hard($R$) for the prover, and publicly easy to verify. We recalled the scrypt construction and a new scheme we call Diodon – weakly MHF with easy verification with secrets. We have shown Equihash – a memory and computation hard PoW based on generalized birthday and shown its relation to VDFs and proofs of sequential work.

## 3.4 Trusted Execution Environments

*Mic Bowman (Intel – Hillsboro, US), transcription from abstract book*



Hardware-based trusted execution has the potential to dramatically improve the scale, efficiency and performance of decentralized applications. There are, however, two positions that are often taken with respect to trusted execution environments (TEE). The first is that the TEE works perfectly. If that were the case, then large portions of cryptographic research would become irrelevant. We know that TEEs (and any other system component) can be attacked so this idealistic assumption is unreasonable. At the other end we could assume that because it can be attacked, a TEE is useless and be ignored. That position is also because it ignores the difficulty in attacking a TEE. A more appropriate view is that the TEE is part of a larger security context. This approach allows for some performance and efficiency improvements and still preserves the overall system security objectives.

## 3.5 Asymmetric Trust

*Christian Cachin (IBM Research – Zürich, CH)*


**Joint work of** Björn Tackmann, Christian Cachin

The Ripple and Stellar blockchain consensus protocols aim at relaxing the strict assumptions in classical consensus and develop so-called federated consensus methods. They intend to stand between traditional BFT consensus (in the sense that the set of nodes is known and pre-agreed) and decentralized consensus (where participation is completely open to anyone). In practice this means that every node declares a list of other nodes which it "trusts." Consensus decisions can be federated in this way, from groups of participants potentially unknown to each other that may have different trust assumptions. For example, one might first reach consensus in small subsets and subsequently combine those partial results in later protocol steps, to reach consensus across the complete system.

Protocols that aim at this goal are not understood well, even though some are used in live blockchains. For example, a recent paper by authors from Ripple casts doubts on the claimed properties of the Ripple protocol. We sketch a model for asymmetric Byzantine quorum systems that explains such protocols precisely. It strictly generalizes existing Byzantine quorum systems. Well-known consensus protocols for Byzantine quorum systems can easily be extended to work in this new model with asymmetric trust.

## 3.6 PERUN – Virtual Payment and State Channels

*Lisa Eckey (TU Darmstadt, DE)*


**Joint work of** Stefan Dziembowski, Lisa Eckey, Sebastian Faust, Daniel Malinowski
**Main reference** Stefan Dziembowski, Lisa Eckey, Sebastian Faust, Daniel Malinowski: "PERUN: Virtual Payment Channels over Cryptographic Currencies", IACR Cryptology ePrint Archive, Vol. 2017, p. 635, 2017.
**URL** http://eprint.iacr.org/2017/635

One approach to securely scale blockchain protocols is to move some of the transaction load off-chain. Payment and State channels allow secure, optimistic, off-chain execution of multiple transactions and even contracts, while only relying on two on-chain interactions with a smart contract. As long as all connected parties agree to the current state of the channel, this method allows for fast and cheap off-chain execution of state changes. The overall security of each direct state channels is guaranteed through a single on-chain contract. Additionally, two existing state channels can be composed to form a new "virtual" state channel, that can be opened and closed off-chain and indirectly connects two parties through a network of state channels. In case of disputes, virtual state channels can be resolved in two ways, either the parties use the connecting intermediary or they directly go to the blockchain. The former solution adds a layer of protection for honest parties since they might not have to deal with on-chain disputes themselves, but it adds risks to the intermediary who might be forced to pay a high amount of transaction fees. Therefore the direct resolve offers a more fair way to resolve disagreement.

## 3.7 Proof of Personhood

*Bryan Ford (EPFL Lausanne, CH), transcription from abstract book*


**Joint work of** Maria Borge, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, Bryan Ford
**Main reference** Maria Borge, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, Bryan Ford: "Proof-of-Personhood: Redemocratizing Permissionless Cryptocurrencies", in Proc. of the 2017 IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2017, Paris, France, April 26-28, 2017, pp. 23–26, IEEE, 2017.
**URL** http://dx.doi.org/10.1109/EuroSPW.2017.46

Decentralization and membership foundations for permissionless blockchains such as proof-of-work, -stake, -storage, -elapsed -time, etc. are all "proof-of-investment": anyone who can and is willing to invest more gets more voting power and rewards. All proof-of-investment schemes are subject to re-centralization. ("rich get richer") due to economies of scale, and cannot provide a human-centric notion of fairness or equity. We propose Proof-of-Personhood, a democratic foundation for blockchain decentralization that gives each (real) person one equal "vote" or unit of "stake." Proof-of-Personhood (PoP) can be implemented in principle using government-issued identities, social trust networks, biometrics or physical presence tests such as pseudonym parties or PoP parties. We are developing tools and processes for PoP parties because they can be made privacy-preserving, low-cost, implementable anywhere including paperless (undocumented) refugees and migrants, and can potentially enforce strong and transparent "one-per-person" Sybil attack resistance. An "anytrust" group of organizers,

secures each party locally, while an inter-party federation or trust network with evidence-based transparency and cross-witnessing processes secures the system globally against corrupt parties. Applications include "one-per-person" accountable anonymous Web browsing or website login tokens, membership tokens for online reputation systems or social media forums, voting tokens for online deliberative forums such as liquid democracy, and one-per-person cryptocurrency minting tokens for decentralized implementations of "Universal Basic Income."

## 3.8 Manipulating Incentives

*Aljosha Judmayer (Secure Business Austria Research, AT)*



The theoretical possibility of bribing attacks on cryptocurrencies has been known since 2015/2016, with various techniques proposed since. The majority of these proposals focus on in-band bribing attacks, executed within the same cryptocurrency they are designed to attack.

This talk presents unpublished research on out-of-band bribing attacks, which are capable of facilitating double-spend collusion across different blockchains.

The bribing logic is thereby managed by a smart contract on a funding cryptocurrency, which leverages cross-chain state verification of the target cryptocurrency to determine the attack outcome and react accordingly.

Contrary to existing schemes, colluding miners are reimbursed independently of success or failure of the attack. This allows our bribing attack to become cheaper than comparable bribing attacks (i.e., the whale attack).

Finally, to hinder counter bribing measures and further reduce the costs of the attack, the notion of crowdfunded bribing attacks is introduced, where the interests of several attackers are aligned by the smart contract to execute multiple double-spending attacks concurrently.

## 3.9 Redesigning Bitcoin's Fee Market

*Ron Lavi (Technion – Haifa, IL)*


**Joint work of** Ron Lavi, Or Sattath, Aviv Zohar
**Main reference** Ron Lavi, Or Sattath, Aviv Zohar: "Redesigning Bitcoin's fee market", CoRR, Vol. abs/1709.08881, 2017.
**URL** https://arxiv.org/abs/1709.08881

Two of Bitcoin's challenges are (i) securing sufficient miner revenues as block rewards decrease, and (ii) alleviating the throughput limitation due to a small maximal block size cap. These issues are strongly related as increasing the maximal block size may decrease revenue due to Bitcoin's pay-your-bid approach. To decouple them we analyze the "monopolistic auction" [Goldberg et al. 2006], showing: (i) its revenue does not decrease as the maximal block size increases, (ii) it is resilient to an untrusted auctioneer (the miner), and (iii) simplicity for transaction issuers (bidders), as the average gain from strategic bid shading (relative to bidding one's true maximal willingness to pay) diminishes as the number of bids increases.

## 3.10 What Can Blockchains Do for You?

*Ian Miers (Cornell Tech – New York, US), transcription from abstract book*



Blockchains are a form of limited trusted third party. Unlike many proposed schemes, they are deployed and readily available. In addition to analyzing how blockchains can be improved by computer science, we should ask what they can do to solve issues in e.g. computer security and cryptography. This talk explored how they can be used to achieve fairness in multi party computation and get general secure computation with state keeping, proof and publications, and assured F/O.

## 3.11 Incentive-Compatibility – A Brief Tutorial

*Tim Roughgarden (Stanford University, US), summarized by Patrik Keller*



Tim Roughgarden gave an introduction to game theory and its application to security. He highlighted the importance of incentive-compatibility and how it can be achieved. Relevant tools for showing incentive-compatibility are dominant strategies and equilibria. Tim explained these concepts in the context of auctions. For auctions, the goal is that all participants truthfully state their willingness to pay. He showed that this is indeed a dominant strategy when participating in a Vickrey auction.

## 3.12 Biologically-Inspired Scaling for Cryptocurrencies

*Marie Vasek (University of New Mexico, US), transcription from abstract book*



All organisms scale in similar ways. For example, the metabolism rate scales at a rate approximately to the 3/4 power. However, information systems in organisms scale differently. We use the immune system, a partially decentralized network, as inspiration for how cryptocurrencies can scale. For instance, all T cells are approved initially by a centralized authority, the thymus. Afterwards, T cells respond relatively decentralized to infection. We outline how cryptocurrencies have many centralized checkpoints and discuss scaling them and the inherent issues therein.

## 3.13 Consensus without Cryptography

*Roger Wattenhofer (ETH Zürich, CH)*



Distributed protocols often employ some form of cryptography, in particular digital signatures. This talk shed some light on the role of cryptography in distributed systems, in particular byzantine agreement (consensus). It presented the simple Ben-Or voting framework, and then discussed various versions of how to implement the random choice: By a local coin, by an oracle, by a pre-determined bit-string. The bit-string must be hidden with Shamir's secret sharing, or even worst-case scheduling will break the framework. In the end, the talk discussed recent alternative methods, in particular the idea to use at least a quadratic number of random bits.

This brings us the interesting question to what degree cryptography is needed: What distributed problems can or cannot be solved without cryptography, and what is still unknown?

# 4 Working Groups

Each of the talks was followed by a short discussion. The discussions which had to be aborted in order to stay in schedule were later continued in smaller break-out groups. We had working groups to the following topics:

- Incentives – Micro to Macro
- Crypto Agility
- Network Layer
- Hardware
- Asymmetric Trust
- Proof-of-X
- Privacy
- Responsible Disclosure
- Game Theory
- Transaction Fees
- Governance
- STARKs

Some of these smaller session were very fruitful while others came to end early. We thus provide abstracts for only some of the working groups.

## 4.1 Crypto Agility

*Patrik Keller (Universität Innsbruck, AT)*



The Crypto Agility session was mainly about protocol migration strategies in case of new weaknesses in the deployed implementations of cryptographic primitives. We started with enumerating the different types of cryptographic breaks:

1. Signature schemes: existential forgery, universal forgery, key recovery
2. Hash functions: collision, preimage recovery, bias towards lower hashes
3. Zero knowledge: soundness failure, failures in zero knowledge

While some breakages seem to be fatal (key recovery, reproducible hash collisions), a carefully crafted protocol may be resilient to a relevant subset of failures. We discussed three strategies to defend against or recover from breaks:

1. Key updates: voluntary or mandatory, gradual update to a different signature scheme if existing scheme is imminently broken.
2. Hybrid accounts: multiple signature schemes in parallel, moving funds requires signatures in all schemes.
3. Backup keys: global activation of a previously set up fallback signature scheme might help in case of a rapidly developing signature scheme break.

## 4.2 Asymmetric Trust

*Patrik Keller (Universität Innsbruck, AT)*



This session followed up on Christian Cachin's talk on Asymmetric Trust. We discussed the following three points:

1. The presented results are based on a predisposed and bounded set of nodes. Can this be adapted for a potentially unlimited number of nodes? Can the results be reused for the permissionless setting, where nodes can join and leave the network at any time?
2. We clarified that the presentation considered only the safety properties of asymmetric trust.
3. We attempted to relate the new notion of Asymmetric Trust to existing work on DAG protocols. Can it be used to obtain improved security proofs?

## 4.3 Proof-of-X

*Patrik Keller (Universität Innsbruck, AT)*



There are two kind of proof-of-X schemes. On the one hand, there are proof-of-consumption schemes where a resource is consumed in order to obtain the right to participate in the protocol. The scarcity of the consumed resource implies a rate limit on participation. On the other hand, there are proof-of-stake schemes where the ownership of a resource is demonstrated instead. The fundamental difference between consumption and ownership makes the two kinds hard to compare. Proof-of-consumption and proof-of-stake must thus be treated separately.

On the proof-of-consumption side, we further recapitulated the properties necessary for proof-of-work constructions as presented by Alex Biryukov.

## 4.4 Responsible Disclosure

*Rainer Böhme (Universität Innsbruck, AT)*



Responsible disclosure of security vulnerabilities poses specific challenges in the domain of cryptocurrencies. The group tried to understand the differences between the conventional debate for proprietary software, which often encompasses national security interests and the specifics of the cryptocurrency space. The participants found similarities (e.g., the absence of a single point of contact also applies to many open source projects), differences in severity (e.g., competition between cryptocurrency projects and the fact that some bugs are easily monetizable), and differences in quality (e.g., some bugs are "unfixable"). The group also compiled a list of issues and plans to write up the lessons learned from case studies in a joint publication.

## 4.5 Transaction Fees

*Patrik Keller (Universität Innsbruck, AT)*



This session arose from the presentation of a new model for Bitcoin transaction fees by Ron Lavi. The presented scheme is based on the assumption of equally sized transactions. We discussed how the new pricing scheme can be adapted to variable transaction size. Additionally we added the constraint of a fixed block size. Unfortunately, the originally tractable optimization problem used for fixing the fee becomes intractable under the additional assumptions.

Apart from that, we observed that cryptocurrencies in contrary to earlier discussed systems (e.g. music download pricing) allow for asymmetric auctions where the buyer pays more than the seller receives. The difference could be burned. Whether this additional freedom in the design space allows for better auction schemes is to be discussed.

## 4.6 Governance

*Assimakis Kattis (New York University, US)*



Governance issues around distributed systems centered around the various potential desirable properties that governance models should have. The main areas of discussion revolved around:

1. Separation of powers between stakeholders
2. Responsiveness to emergencies
3. Transparency in the governance process
4. Accountability of actors to each other and the general ecosystem

Furthermore, the techniques to ensure decentralization and a reliable practical implementation of governance goals were identified as important areas for further study. Interaction with the current legal system was also discussed, along with the questions it poses for the design of governance models.

## 4.7 STARKs

*Assimakis Kattis (New York University, US)*



This session was based around Eli Ben-Sasson's talk on STARKs and their potential for scalability. We looked at the theoretical foundations of STARKs and investigated what the main barriers for scalable implementations are. Discussions followed around STARK trust assumptions, proof sizes and use cases. The separation in prover efficiency between non-interactive protocols and PCPs/IOPs was analyzed, along with efficient constructions of the latter.

Mathematical tools that allow for efficient STARK representations were also investigated. Usage of low-degree extensions, the relationship between error and soundness, and FFTs for interpolation/evaluation of low-degree polynomials were the main topics covered around the design techniques for STARK proofs. In a subsequent session, the construction and guarantees of getting low degree polynomials to verify were discussed, since they are linked to final proof sizes. We also looked at the potential for compression of constraint checks in STARKs, as well as at hash functions with efficient SNARK/STARK representations.

# 5 Open Problems

*Joseph Bonneau (New York University, US)*



We ended the seminar with a brief discussion of open problems. Many arose throughout the breakout groups, but there were a number of topics we did not have time to full explore. Some of the most interesting included:

- **Unifying proof-of-stake with proof-of-work** Efforts in the "Proof-of-X" session to identify a unified model for analyzing proof-of-stake protocols with proof-of-work protocols did not succeed. Can a unified model be found?
- **Auction-based consensus protocols** There were some efforts to outline a consensus protocol based on miners bidding for the right to create a block. The idea seems promising but we could not agree on an exact protocol.
- **On-chain governance** Most of the discussion of governance focused on higher-level decision making about protocol governance. It is an interesting question to explore what sorts of governance decisions can be made automatically by voting on the chain itself.
- **Incorporating market impacts** It is widely agreed that game-theoretic models of cryptocurrency should eventually incorporate the notion of market impact: executing attack may affect exchange rates and hence hurt the attacker despite a nominal gain in rewards. We lack a clear roadmap for incorporating this phenomenon into models in a tractable way.

## Participants


- Svetlana Abramova
Universität Innsbruck, AT
- Sarah Azouvi
University College London, GB
- Foteini Baldimtsi
George Mason University – Fairfax, US
- Eli Ben-Sasson
Technion – Haifa, IL
- Alex Biryukov
University of Luxembourg, LU
- Rainer Böhme
Universität Innsbruck, AT
- Joseph Bonneau
New York University, US
- Mic Bowman
Intel – Hillsboro, US
- Dominic Breuker
solarisBank AG, DE
- Christian Cachin
IBM Research-Zurich, CH
- Nicolas Christin
Carnegie Mellon University – Pittsburgh, US
- Lisa Eckey
TU Darmstadt, DE
- Ittay Eyal
Technion – Haifa, IL
- Bryan Ford
EPFL Lausanne, CH
- Christina Garman
Purdue University – West Lafayette, US
- Arthur Gervais
Imperial College London, GB
- Philipp Jovanovic
EPFL Lausanne, CH
- Aljosha Judmayer
Secure Business Austria Research, AT
- Ghassan Karame
NEC Laboratories Europe – Heidelberg, DE
- Assimakis Agamemnon Kattis
New York, US
- Stefan Katzenbeisser
TU Darmstadt, DE
- Patrik Keller
Universität Innsbruck, AT
- Ron Lavi
Technion – Haifa, IL
- Patrick McCorry
King's College London, GB
- Ian Miers
Cornell Tech – New York, US
- Malte Möser
Princeton University, US
- Tyler W. Moore
University of Tulsa, US
- Neha Narula
MIT – Cambridge, US
- Tim Roughgarden
Stanford University, US
- Tim Ruffing
Universität des Saarlandes, DE
- Emin Gün Sirer
Cornell University – Ithaca, US
- Yonatan Sompolinsky
The Hebrew University of Jerusalem, IL
- Itay Tsabary
Technion – Haifa, IL
- Florian Tschorsch
TU Berlin, DE
- Marie Vasek
University of New Mexico, US
- Roger Wattenhofer
ETH Zürich, CH
- Edgar Weippl
Secure Business Austria Research, AT
- Aviv Zohar
The Hebrew University of Jerusalem, IL


Report from Dagstuhl Seminar 18462

# Provenance and Logging for Sense Making

**Edited by**

**Jean-Daniel Fekete[1], T. J. Jankun-Kelly[2], Melanie Tory[3], and Kai Xu[4]**

1 **INRIA Saclay – Orsay, FR,** jean-daniel.fekete@inria.fr
2 **Mississippi State University, US,** tjk@acm.org
3 **Tableau Software – Palo Alto, US,** mtory@tableau.com
4 **Middlesex University – London, GB,** k.xu@mdx.ac.uk

## Abstract

Sense making is one of the biggest challenges in data analysis faced by both the industry and the research community. It involves understanding the data and uncovering its model, generating a hypothesis, selecting analysis methods, creating novel solutions, designing evaluation, and also critical thinking and learning wherever needed. The research and development for such sense making tasks lags far behind the fast-changing user needs, such as those that emerged recently as the result of so-called "Big Data". As a result, sense making is often performed manually and the limited human cognition capability becomes the bottleneck of sense making in data analysis and decision making.

One of the recent advances in sense making research is the capture, visualization, and analysis of provenance information. Provenance is the history and context of sense making, including the data/analysis used and the users' critical thinking process. It has been shown that provenance can effectively support many sense making tasks. For instance, provenance can provide an overview of what has been examined and reveal gaps like unexplored information or solution possibilities. Besides, provenance can support collaborative sense making and communication by sharing the rich context of the sense making process.

Besides data analysis and decision making, provenance has been studied in many other fields, sometimes under different names, for different types of sense making. For example, the Human-Computer Interaction community relies on the analysis of logging to understand user behaviors and intentions; the WWW and database community has been working on data lineage to understand uncertainty and trustworthiness; and finally, reproducible science heavily relies on provenance to improve the reliability and efficiency of scientific research.

This Dagstuhl Seminar brought together researchers from the diverse fields that relate to provenance and sense making to foster cross-community collaboration. Shared challenges were identified and progress has been made towards developing novel solutions.

**Seminar** November 11–16, 2018 – http://www.dagstuhl.de/18462
**2012 ACM Subject Classification** Human-centered computing → Visualization theory, concepts and paradigms
**Keywords and phrases** Logging, Provenance, Sensemaking, Visualization
**Digital Object Identifier** 10.4230/DagRep.8.11.35
**Edited in cooperation with** Christian Bors


Provenance and Logging for Sense Making, *Dagstuhl Reports*, Vol. 8, Issue 11, pp. 35–62
Editors: Jean-Daniel Fekete, T. J. Jankun-Kelly, Melanie Tory, and Kai Xu

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

# 1 Executive Summary

*Kai Xu (Middlesex University – London, GB)*
*Jean-Daniel Fekete (INRIA Saclay – Orsay, FR)*
*T. J. Jankun-Kelly (Mississippi State University, US)*
*Melanie Tory (Tableau Software – Palo Alto, US)*



Sense making is one of the biggest challenges in data analysis faced by both industry and research community. It involves understanding the data and uncovering its model, generating hypothesis, select analysis methods, creating novel solutions, designing evaluation, and the critical thinking and learning wherever needed. Recently many techniques and software tools have become available to address the challenges of so-called 'Big Data'. However, these mostly target lower-level sense making tasks such as storage and search. There is limited support for the higher-level sense making tasks mentioned earlier. As a result, these tasks are often performed manually and the limited human cognition capability becomes the bottleneck, negatively impacting data analysis and decision making. This applies to both industry and academia. Scientific research is a sense making process as well: it includes all the sense making tasks mentioned earlier, with an emphasis on the generation of novel solutions. Similar to data analysis, most of these are conducted manually and considerably limit the progress of scientific discovery.

*Visual Analytics* is a fast-growing field that specifically targets sense making [6]. It achieves this by integrating interactive visualization with data analytics such as *Machine Learning*. It follows a human-centered principle: instead of replacing human thinking and expertise with algorithms and models, it enables the two to work together to achieve the best sense making result. Fast progress has been made in the last decade or so, which is evidenced by the publications in the Visual Analytics conferences such as IEEE VAST (part of IEEE VIS) and the increasing popularity of visual approaches in many other fields such as Machine Learning, Information Retrieval, and Databases.

One recent advance in Visual Analytics research is the capture, visualization, and analysis of *provenance* information. Provenance is the history and context of sense making, including the "7W" (Who, When, What, Why, Where, Which, and HoW) of data used and the users' critical thinking process. The concept of provenance is not entirely new. In 1996, Shneiderman recognized the importance of provenance by classifying *history* as one of the seven fundamental tasks in data visualization [4]. History allows users to review previous actions during visual exploration, which is typically long and complex. Provenance can provide an overview of what has been examined and reveal the gap of unexplored data or solutions. Provenance can also support collaborative sense making and communication by sharing the rich context of what others have accomplished [7].

The topic of provenance has been studied in many other fields, such as Human-Computer Interaction (HCI), WWW, Database, and Reproducible Science. The HCI research community heavily relies on user information, such as logging and observation, in their study. These closely relate to provenance and share the common goal of making sense of user behavior and thinking. The collaboration between the two fields can potential create novel solutions for some long-standing research challenges. For instance, it has been shown that provenance information can be used to semi-automate part of the qualitative analysis of user evaluation data [3], which is notoriously time-consuming.

The WWW and Database research community has been actively working on provenance for the last decade or so, with a particular focus on tracking data collection and processing.

This has led to the recent publication of the W3C reference model on provenance [1]. A important part of these efforts is to make sense of the source and quality of the data and the analyses base on them, which has a significant impact on their uncertainty and trustworthiness [1]. Similarly, there is a fast growing Reproducible Science community, whose interest in provenance is "improving the reliability and efficiency of scientific research ... increase the credibility of the published scientific literature and accelerate discovery" [2].

There is a trend of cross-community collaboration on provenance-related research, which has led to some exciting outcomes such as the work integrating visualization with reproducible science [5, 8]. However, there are still many challenging research questions and many provenance-related research efforts remain disconnected. This seminar brought together researchers from the diverse fields that relate to provenance. Shared challenges were identified and progress has been made towards developing novel solutions.

The main research question that this seminar aims to address is: **How to collect, analyze, and summarize provenance information to support the design and evaluation of novel techniques for sense making across related fields**. The week-long seminar started with a day of self-introduction, lighting talks, and research topic brain storming. The self-introduction allowed attendees to know each other better, and the lighting talks covered the latest work in the research fields related to provenance. Each participant proposed several research questions, which were then collated and voted on to form the breakout groups. The following are the research areas chosen by the participants:

- Storytelling and narrative;
- Provenance standard and system integration;
- Task abstraction for provenance analysis;
- Machine learning and provenance;
- User modeling and intent.

The rest of the week was breakout session, and each participant had the option to change group halfway. The seminar finished with a presentation from each group and discussions on the next steps to continue the collaboration. Many interesting problems were identified, and progress was made towards new solutions. Please refer to the rest of the report for the details on the identified research questions and the progress made by the end of week.

## References


1 Melanie Herschel and Marcel Hlawatsch. Provenance: On and Behind the Screens. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 2213–2217, New York, NY, USA, 2016. ACM.

2 Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, January 2017.

3 P. H. Nguyen, K. Xu, A. Wheat, B. L. W. Wong, S. Attfield, and B. Fields. SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):41–50, January 2016.

4 Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

5 C. T. Silva, J. Freire, and S. P. Callahan. Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science Engineering*, 9(5):82–89, September 2007.


[1] https://www.w3.org/TR/prov-overview/

6 James J. Thomas and Kristin A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Centre, 2005.

7 Kai Xu, Simon Attfield, T. J. Jankun-Kelly, Ashley Wheat, Phong H. Nguyen, and Nallini Selvaraj. Analytic provenance for sensemaking: A research agenda. *IEEE Computer Graphics and Applications*, 35(3):56–64, 2015.

8 Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. Controlling false discoveries during interactive data exploration. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 527–540, New York, NY, USA, 2017. ACM.

## 2 Table of Contents

# 3 Overview of Talks

## 3.1 Provenance in Databases

*Leilani Battle (University of Maryland – College Park, US)*



Recording the history (or provenance) of how a dataset was processed and making the history accessible through an intuitive interface is a challenging problem. From a database management systems (DBMS) perspective, the focus is on performance: efficiently recording information about how database queries process input data, enabling fast analysis of the results via database queries over the provenance data, and exploiting recorded provenance in downstream applications connected to the DBMS. In this presentation, I discuss the data management motivations, challenges and techniques for supporting provenance-based analysis.

## 3.2 Sensemaking, and the Analytic Provenance of Sense

*Alex Endert (Georgia Institute of Technology – Atlanta, US)*



People perform exploratory data analysis to gain insight and make sense of their data. This process is often called "sensemaking", and conceptual models exist that help describe the formation of mental models. The process of performing sensemaking often involves many operations, tasks, and interactions. "Analytic Provenance" is a term that describes the study of this iterative interactive process. This talk will introduce these two concepts, how they relate to each other, and what primary research challenges are in each of them.

## 3.3 (A Blazingly Fast Intro to) Reproducible Science

*Claudio T. Silva (New York University, US)*



In this ten minute talk, we give a short introduction to reproducible science. Using two examples of our research, we show the success and failure of producing results that can be reproduced. Our experience is similar to others, as noted in the reproducibility survey done by Nature [1]. We use as examples the recently established ACM Artifact Review and Badging policy, the ACM SIGMOD reproducibility effort, and the Graphics Replicability Stamp Initiative to highlight issues related to repeatability, replicability, and reproducibility of scientific results. We end with a brief discussion of the "reproducible paper", where we use the VisTrails system to propose a provenance-based infrastructure to support the life cycle of papers [2].

**References**

1 Monya Baker. Is there a reproducibility crisis? a nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help. *Nature*, 533(7604):452–455, 2016.

2 David Koop, Emanuele Santos, Phillip Mates, Huy T Vo, Philippe Bonnet, Bela Bauer, Brigitte Surer, Matthias Troyer, Dean N Williams, Joel E Tohline, et al. A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Computer Science*, 4:648–657, 2011.

## 3.4 Provenance and Logging: HCI Perspectives

*Melanie Tory (Tableau Software – Palo Alto, US)*



This talk explores provenance and logging from the perspective of two distinct user groups: data analysts and researchers. Analysts need provenance information to support their analytical workflow, including review and reflection, collaboration, learning, and storytelling. HCI researchers need provenance information to understand user behavior and infer barriers to workflow where design changes may be warranted. In both cases, I argue that raw provenance information is not enough; it may need to be transformed and summarized to support user needs.

# 4 Working groups

## 4.1 Storytelling

*Daniel Archambault (Swansea University, GB), T. J. Jankun-Kelly (Mississippi State University, US), Andreas Kerren (Linnaeus University – Växjö, SE), Robert Kosara (Tableau Software – Seattle, US), Ali Sarvghad (University of Massachusetts – Amherst, US), and William Wong (Middlesex University – London, GB)*



In the context of storytelling, trust is how confident the audience/author is of the story contents. Depending on the audience, more or less provenance information would be required to establish trust. The first step in establishing trust is establishing authority. One way to establish authority is to demonstrate the provenance of the story to the audience. In particular, we show the data and the provenance information of how we got there. If both of these are available in a complete way to the audience, this can be one way to convince the audience that the author is trustworthy. A transparent communication of this information allows the audience to verify and replicate the steps leading to the creation of this story.

Counterstories are stories that run contrary or lead to the opposite conclusion of the proposed story given the data. With interaction and data provenance information, counterstory development can be supported. If during the presentation of a story, the author can convey that the counterstories are less probable than the proposed story, this can provide evidence that the proposed story is more trustworthy. Conversely, if an author finds the counterstory more credible, it can be promoted to the proposed story.

An ideal system would be able to propose areas of the data set that have not been explored in an intelligent way to verify that the exploration is complete. This process can be partially automated by metrics on the data set and can be influenced by interactive pruning of the provenance tree – the user annotates sections of the tree as uninteresting, providing some text to state why it is not relevant to the exploration. Thus, the creation of a credible story could follow this procedure:

**Auto suggestions → Story ← User notes on analysis**

For completeness, automated tools suggests areas of underexplored data guided by user annotations on the provenance tree to indicate why certain branches are not relevant to this analysis. The process converges on a credible story.

## 4.2 VAPS: Visual Analytics Provenance Standard for Cross-Tool Integration of Provenance Handling

*Daniel Archambault (Swansea University, GB), Jean-Daniel Fekete (INRIA Saclay – Orsay, FR), Melanie Herschel (Universität Stuttgart, DE), T. J. Jankun-Kelly (Mississippi State University, US), Andreas Kerren (Linnaeus University – Växjö, SE), Robert S. Laramee (Swansea University, GB), Aran Lunzer (OS Vision – Los Angeles, US), Holger Stitz (Johannes Kepler Universität Linz, AT), and Melanie Tory (Tableau Software – Palo Alto, US)*



The state of the art in capture and presentation of provenance for visual analytics is exemplified by a number of systems that support activities within a single tool, or at best across a set of tools that are constrained to work within a single predefined pattern of activity. For example, VisTrails [1], while equipped with provenance management and a provenance editor that support a class of workflow-style activities, cannot collect provenance outside its own world. However, visual analytics activities seldom take place within a single tool; it is typical for an analyst to refer to, and combine findings from, several tools that are running in parallel. These may include data wrangling programs, databases, visual analytics tools, presentation tools, and reporting systems. Managing provenance from within only one of these tools is hiding a large portion of the analytical work and—more importantly—opportunities to make sense of the user's activity through the observation of the provenance data. For example, a user taking a screen capture of the visualization application and pasting it into a presentation tool is an obvious sign that an insight (or a bug) has been discovered, but the visualization application has no way to know that a screen capture has been performed and will not see that insight event. Therefore, we believe a provenance management system should handle provenance data from multiple applications, and also show inter-application activities (e.g., cut/paste, screen-capture/paste). To achieve this vision, we should address multiple problems that are described in the next section. Some are challenging, and the challenges are listed and discussed in the second section. Fortunately, we envision a possible solution that we discuss in the third section. It will not solve the problem completely and will need time to improve, but we believe that existing technologies can be used to address the main issues in a realistic way.

### 4.2.1 Problem Statement

Managing provenance from multiple applications requires that all the applications record their provenance in a way that can be interpreted by the provenance visualization and management tool (PVMT) that is being used to oversee the analysis session. This suggests the need for a standard format for provenance data. While no such format exists at present, the striking similarity between the trace formats used by existing applications leads us to believe that a concrete standard format can be specified with little controversy.

In addition, to be useful, a concrete trace format should present activity information at several levels of precision/granularity. The raw format is generated directly by the tool using its internal actions and parameters, but a more usable format would used abstract actions understandable regardless of the actual tool, such as "select", "pan", and "zoom". We therefore need to create a vocabulary of these abstract actions that all the PVMT(s) will recognize and visualize in an understandable way. Among these abstract actions, some virtual actions should also be defined. For example, it should be possible to annotate the provenance trace with screenshots to keep track of the visible state of the program at some key points. These annotations can be used by the PVMT to help navigate the provenance tree.

Managing provenance at a higher level also requires the PVMT to be able to talk back to applications and control them to navigate the provenance graph for undo/redo, or more generally jump to multiple steps of the graph. Therefore, in addition to a standard format, we need a standard service to manage the communication between program generating provenance traces and the PVMT(s). Again, this service will require some new information in the provenance trace. For example, the "undo" function returns to a particular state in the past, which may encompass several actions from the provenance trace. Marking the scope of "atomic" actions in the trace is essential to visualize a meaningful chunking of states that can be targets for navigation.

It is not yet clear if one level of abstraction is sufficient, but it is clear that a PVMT designed for debugging will want to show low-level trace actions, whereas one dedicated to support users might only show abstract ones. Summaries of large amounts of traces might need an event more abstract level; this will be investigated later.

### 4.2.2 Related Work

There has been substantial work in the domain of provenance capture, visualization, analysis, and reuse for visualization and visual analytics. VisTrails [1] showed the way with an integrated environment in Python, able to run complex analytics pipelines, keeping track of their development over time, editing the history of the evolution of the pipeline as a tree of pipelines evolutions, and allowing analysis to navigate in the provenance graph, replaying or continuing analyses. Other systems such as CZSaw [2] have built upon the experience of VisTrails to let graph visualization and analysis benefit from provenance management. More recently, Caleydo [4] has started to offer provenance capture and management for Web-based analytics systems. However, the landscape of data analytics has evolved and it is clear that analysts are always working with multiple applications and instrumenting only one will not capture the whole analytics process.

On the other side, there is a whole community working on provenance management. This community is related to Databases and Operating Systems, and has been running the International Provenance and Annotation Workshop (IPAW) [3] since 2002, as well as the "Workshop on the Theory and Practice of Provenance" (TAPP) Usenix conference until 2009.

The PROV standard has been published by the W3C[2] and later the ProvONE standard more suited for Scientific Dataflow systems[3].

Meanwhile, the visualization and visual analytics community have published many articles explaining systems and techniques for managing provenance and visualizing it. The LIVVIL workshop, organized at VIS 2014[4], has been the inspiration of the Dagstuhl seminar on Provenance and Logging for Sensemaking[5].

However, despite all these work related to provenance, the data analysis community is still left without practical solutions to match the promises showcased by VisTrails two decades ago. The community should get inspiration from all the research and experiments done so far to come out with a solution to manage provenance for data analysis, leading to sensemaking, storytelling, and hopefully many other outcomes demonstrated by VisTrails and its long list of related publications. Challenges There are a number of challenges involved in building such a system. We focus on the principal challenges involved in a standardized file format for logging provenance in a tool-agnostic way. Such a format, would need to capture the following in a scalable way: high-level events and low-level events that are common to many tools. The low-level events could be implemented in a tool-agnostic way whereas the high-level events would be application specific and would need to be defined in a way that extends the standard. There is the potential to not only define these events but provide a classification of such events. The standard should be easy to implement and extend so that new visualization tools can implement this standard.

### 4.2.3 Possible Solution and Opportunities

Transform RAW provenance to provenance standard Might be provided by the tool author In the future tools can directly provide provenance information that complies to the standard Provide a service that manages the standardized provenance Store/retrieve provenance information Execute actions from the provenance graph on application Result are multiple provenance trees from different applications that can be combined Provides a holistic view on the analysis process and might offer better insights Hence, new visualization approaches to mine the provenance information are required

### Tasks and Users of the Provenance

The suggested approaches and capabilities were informed by a series of tasks and a group of three roles (i.e., users, designers, others) that could require those tasks. These tasks were: Remembering where one was in the exploration (primarily for users), explaining the exploration and findings for others (by the users for others), navigation during the exploration (users), meta-analysis of the exploration (for designers to see how the tool is used or others to see many traces), debugging/optimization of traces (for designers to retool their systems), and for reproduction of the results from the same data at different times (users and others), and using the trace as a template on different data (for users and others). For a deeper exploration of these needs, see Ragan [7].

[2] https://www.w3.org/TR/prov-overview/
[3] http://vcvcomputing.com/provone/provone.html
[4] https://livvil.github.io/workshop/
[5] https://www.dagstuhl.de/18462

### What about the Vis?

While the discussion primarily focused on what is needed to support visualization tools for provenance, some time was given to what types of visualizations could be used. Traditionally, tree-like visualizations have been used to depict the branching trace structure (e.g., VisTrails [1]). However, each of the different roles we touched upon could use novel (or at least, not strictly tree-based) visualizations. Linear temporal sequences provide a clear sense of the direct history of the exploration. EventFlow-like compressed trees [6] lose the temporal aspects, but can highlight patterns. Graphs of the parameter similarity or exploration depth built upon metrics can also be considered [5]. When comparing or trying to apply different trees, means of overlaying or visually querying them are needed. Other novel approaches can be explored given the suggested framework. Extension: What activities could be supported using interactive visualizations? Delivering a visualization of the activities that have been carried out during some analysis session is only part of the story. With the appropriate API-based connections to the applications that generated the provenance records, a visualization can provide the means for a user to revisit the state of the session at any point that has been logged. This would mean, for example, re-establishing the full context of data selections and chart settings in an application such as Tableau, ideally including the full interaction history so that the user is faithfully transported back to the set of decision possibilities – including visualization adjustments, and undo/redo opportunities – that were available to the original analyst in that moment. One way for the developers of an application to support at least a limited form of such state revisiting would be through parameterized URLs, that are cheap to embed in the provenance stream and whose parameters will typically be tailored to the needs of the specific application. The more general solution, of course, would be for the application to accept from the visualization a sub-range of the entire provenance stream from the start of the recorded session up to the point of interest, and to use this stream to rebuild its state. Our hope is that application developers would see it as being in their interest to provide this maximal form of faithful reinstatement.

### References


1 Louis Bavoil, Steven P. Callahan, Carlos Eduardo Scheidegger, Huy T. Vo, Patricia Crossno, Cláudio T. Silva, and Juliana Freire. Vistrails: Enabling interactive multiple-view visualizations. In *16th IEEE Visualization Conference, VIS 2005, Minneapolis, MN, USA, October 23-28, 2005*, pages 135–142. IEEE Computer Society, 2005.

2 Yingjie Victor Chen, Zhenyu Cheryl Qian, Robert Woodbury, John Dill, and Chris D. Shaw. Employing a parametric model for analytic provenance. *ACM Trans. Interact. Intell. Syst.*, 4(1):6:1–6:32, April 2014.

3 Juliana Freire, David Koop, and Luc Moreau, editors. *Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers*, volume 5272 of *Lecture Notes in Computer Science*. Springer, 2008.

4 Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Nicola Cosgrove, and Marc Streit. From visual exploration to storytelling and back again. *Computer Graphics Forum (EuroVis '16)*, 35(3):491–500, jun 2016.

5 T. J. Jankun-Kelly. Using visualization process graphs to improve visualization exploration. In Freire et al. [3], pages 78–91.

6 Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. Temporal event sequence simplification. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2227–2236, 2013.

7 Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Trans. Vis. Comput. Graph.*, 22(1):31–40, 2016.

## 4.3 A Novel Approach to Task Abstraction to Make Better Sense of Provenance Data

*Christian Bors (Technische Universität Wien, AT), Simon Attfield (Middlesex University – London, GB), Leilani Battle (University of Maryland – College Park, US), Michelle Dowling (Virginia Polytechnic Institute – Blacksburg, US), Alex Endert (Georgia Institute of Technology – Atlanta, US), Steffen Koch (Universität Stuttgart, DE), Olga A. Kulyk (DEMCON – Enschede, NL), Robert S. Laramee (Swansea University, GB), Melanie Tory (Tableau Software – Palo Alto, US), and John Wenskovitch (Virginia Polytechnic Institute – Blacksburg, US)*



In any exploratory activity, in which we include sensemaking with Visual Analytics tools [3], emerging and exploratory paths of interaction can lead to new understandings which can be at the same time surprising and unexpected. Indeed, the often inductive, exploratory quality of Visual Analytics is part of its premise. Hence, there are seldom set plans or procedures. In any exploration, it can be important to reconstruct what was done in order to fully interpret an output [1], to judge its validity, perhaps to see what other ground may be covered [4], or simply to learn [2]. Hence, there is value in recording and reconstructing provenance trails. Uninterpreted interaction logs, however, are typically detailed, low-level, and fail to provide ease of overview and rapid insight. Low level provenance data is limited by its lack of an organizing structure and hence a framework with which to make sense of this data. Providing a robust task abstraction framework (or interpretive structure) has the potential to provide the means of using low-level provenance data to construct higher-level task hierarchies (explicit tasks, and relationships between them), allowing users to interpret and gain the benefit from provenance data more easily.

We propose a conceptual task abstraction framework as an approach to enabling meaningful mapping between raw provenance trails and higher-level descriptions of tasks. We assume first a recorded trail of interaction. Next we assume that a task abstraction from this trail can be understood such that higher-level, more abstract task descriptions supervene over lower-level events or actions and provide a shorthand for sequences at the lower-level. Further, we assume that such relationships can be embedded in multiple layers, and hence a multi-leveled hierarchy. One premise of our approach, however, is the claim that abstracted descriptions are both interpreted and dependent on context. In many ways, we take the situated nature of language, and its translation of a series of contextually bound low-level phonemes into a higher-level message which can be further summarized and so on, as a metaphor for the interpretation of provenance data. As a consequence, we do not assume any fixed task hierarchies for a given string of low-level actions, but rely instead on the idea of interpretation as constituted from the construction of ad hoc hierarchies depending on context.

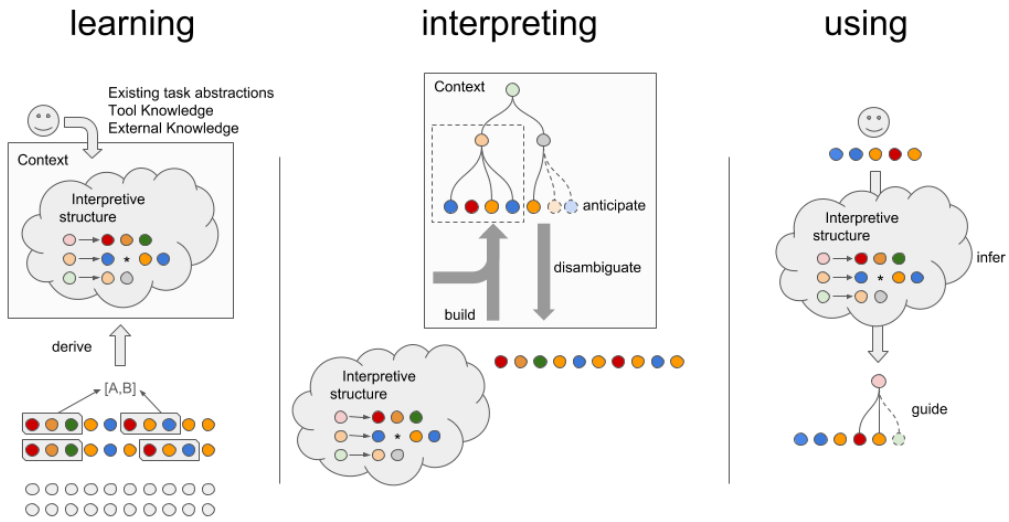



**Figure 1** The three stages for task abstraction based on mapping low level provenance to a higher-level interpretive structure. While one objective of the task abstraction is building a structure for later use, the other is applying this structure for applying it to infer user goals.

### 4.3.1 A Task Abstraction Framework

Our framework can be divided into three major phases: Learn, Interpret, Use. Sequentially, the abstraction will impact these phases differently and changes in any of them will also propagate into the others. Additionally, we exemplify the levels of abstraction in the interpretation of a natural language.

#### 4.3.1.1 Learn

The first phase is the creation of a specific hierarchical structure of *actions*, *tasks*, and *intents* (i.e., an interpretive structure). Such a structure can be created manually, either a priori or in an iterative manner, by deriving it from pre-defined interactions and tasks available in the system that the log traces are recorded from. Alternatively, the structure can be learned/acquired from a (large) number of logged sessions through various means, like sequence mining, process mining, ontology learning, etc.
Following the example of language, an analogy for the learn phase of task abstractions would be determining parts of speech that individual words belong to, based on their use within the context of the language (cf. Figure 2). Consecutively, when learning a language, single words are identified and combined into phrases, determining in which context the phrase can be used.

#### 4.3.1.2 Interpret

The second phase comprises of the interpretation of logged data/provenance information by matching it to available structures determined during the Learn phase. Factors to consider for this are the expressiveness and flexibility of the interactive visualization tools provenance is obtained from, and the peculiarity of the user's task with respect to the concreteness. However, interpretation is also influenced by context that requires a formalization within the framework/structure.

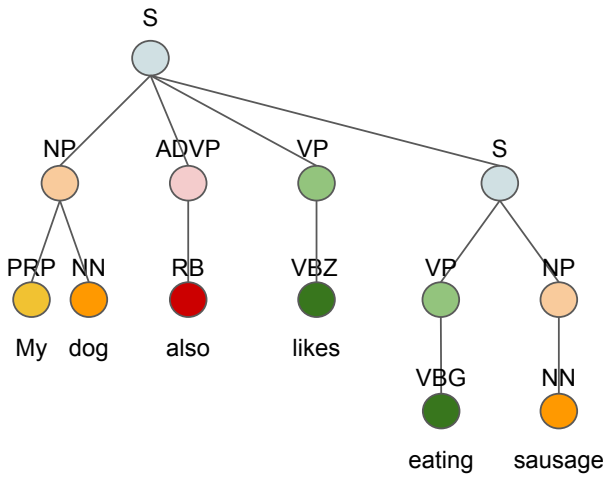



**Figure 2** A parse tree exemplifying our natural language analogy. We parse words into logical phrases. From known phrases we can derive the meaning of unknown words or phrases in the context of the remaining sentence. We can furthermore apply this acquired knowledge to deduce further context (e.g., linking the phrases dog and sausage positively).

Continuing with our natural language analogy, we can use the phrases that we have so far learned to derive the meaning of new vocabulary or phrases that co-occur with known ones.

#### 4.3.1.3 Use

Given a task hierarchy interpreted from an analysis session, there are different uses of the interpretive structure that can be presented to users (e.g., categories of guidance, automatic processing for presentation, ...). Based on the different goals of users, we have to apply the interpretive structure differently to account for the different expectations. For instance, a system could: (a) present templates for how to complete a task; (b) suggest next operations that guide the user to complete a task; (c) provide suggestions on what could be the next step as a decision making guidance; (d) optimize other aspects of the visual data exploration process based on the understanding of a user's process / behavior interpreted from the learned task abstraction; or (e) give an overview to a developer on how users actually conducted a task or tried completing it. Included in these complex UI and UX goals are also challenges with understanding when to present guidance and feedback to users in an appropriate way to minimize interruption and frustration.
Coming back to our example, reading sentences written in our natural language, we can exploit our prior experience with different language conventions and structures to try to deduce the meaning of unfamiliar words and phrases through their positioning, relative to known phrases. The structure, tenses, prepositions of the sentences change, based on how the information is intended to be conveyed.

### 4.3.2 The Role of Context and Uncertainty

Interpretation of user tasks based on low-level interactions is an imperfect process, often with incomplete or uncertain results. For example, some observed provenance records may not "fit" within the system's current interpretation of the user's behavior, and multiple interpretations may appear equally plausible for a given set of provenance records. Furthemore, context can

influence interpretability and omit unlikely outcomes. Context can disambiguate outcomes by narrowing down possibilities based on usage (e.g., differences in individual users' tool expertise), environment variations (e.g., different designs of visualization systems, employed visual encodings), or the application/analysis domain. Additionally, we can incorporate notions of confidence or error into plausible interpretations and update these confidence measures as the user performs more interactions with the system, again drawing additional information from provenance. These measures can then be used to refine the system's understanding of what the user is trying to accomplish over time.

A user may repeat some interaction patterns in their future analysis, but they could also start exhibiting new patterns of analysis. As such, the system should be able to adapt by continuing to refine the interpretive structure given additional input. An important consideration for an active learning approach is the matter of temporal context: when should a system adapt the interface because of learning? When should a system remain the same? Previous work on adaptive menus highlights this problem: Users often rely on consistency to enhance performance, and thus user performance can suffer when consistency is ignored. The Show Me automated presentation feature in Tableau [5] is an example of how consistent recommendations help users to reason about and utilize recommendations efficiently. Alternatively, the uncertainty of outcomes can be actively communicated to the user to clarify expected outcomes and improve accuracy of the interpretive structure.

We are aware of existing research to address aspects of the challenges outlined above, but we see shortcomings in combining them into a singular organizing framework that leverages these approaches. For example, how to manage the uncertainty of interpretation and inherent variability in analysis sessions across user expertise, system design, etc., is currently unclear. Thus, there exist exciting opportunities and challenges along these lines of research that can advance our understanding in how to learn useful structures from provenance. Possible interpretations of our conceptual framework could be implemented by probabilistic parsing, building grammars (NLP), or machine learning (cf. [6, 7, 8, 9, 10]).

## References


1 Bodesinsky, Peter, et al. "Exploration and assessment of event data." Proceedings of EuroVis Workshop on Visual Analytics. 2015.

2 E. T. Brown et al., "Finding Waldo: Learning about Users from their Interactions," in IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1663-1672, 31 Dec. 2014.

3 Pirolli, Peter, and Stuart Card. "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis." Proceedings of International Conference on Intelligence Analysis. Vol. 5. 2005.

4 H. Stitz, S. Gratzl, H. Piringer, T. Zichner, und M. Streit, "KnowledgePearls: Provenance-Based Visualization Retrieval", IEEE Transactions on Visualization and Computer Graphics, p. 1–1, 2018.

5 Tableau Software:. URL: https://www.tableau.com/. Accessed: Jan, 2019.

6 I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques, 4th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016.

7 L. Bradel, C. North, L. House, S. Leman, Multi-model semantic interaction for text analytics, in: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), 2014, pp. 163–172. doi:10.1109/VAST.2014.7042492.

8 J. Z. Self, R. K. Vinayagam, J. T. Fry, C. North, Bridging the gap between user intention and model parameters for human-in-the-loop data analytics, in: Proceedings of the

Workshop on Human-In-the-Loop Data Analytics, HILDA '16, ACM, New York, NY, USA, 2016, pp. 3:1–3:6.doi:10.1145/2939502.2939505.

9 J. Wenskovitch, C. North, Observation-level interaction with clustering and dimension reduction algorithms, in: Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA'17, ACM, New York, NY, USA, 2017, pp. 14:1–14:6. doi:10.1145/3077257.3077259.

10 M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, C. North, Sirius: Dual, symmetric, interactive dimension reductions, IEEE Transactions on Visualization and Computer Graphics 25 (1) (2019) 172–182. doi:10.1109/TVCG.2018.2865047.

## 4.4 Machine Learning and Provenance in Visual Analytics

*Christopher Collins (UOIT – Oshawa, CA), Sara Alspaugh (Splunk Inc. – San Francisco, US), Remco Chang (Tufts University – Medford, US), Phong H. Nguyen (City – University of London, GB), and Kai Xu (Middlesex University – London, GB)*



Provenance and machine learning can be mutually beneficial in two distinct ways: applying machine learning on provenance data to carry out tasks, and using provenance data from machine learning processes to improve the machine learning systems.

In both of these scenarios, the problem of chunking or segmenting event logs is relevant. In particular, we are interested in the opportunities for using interaction records, or logs of user actions, in analytic systems. Interaction records are different from other types of events that a system may record, such as receipt of data from an asynchronous process, or exceptions in network communication. Interaction records relate to task analysis in visual analytics in that a series of low-level interactions may represent a higher level analytic process [5]. For example, hovering on an item, panning the screen, then hovering another item may represent an explore action. We are interested in determining the appropriate chunking of low-level actions, such that they can be used in the variety of applications described below.

### 4.4.1 Machine Learning on Provenance Data

Applications of machine learning on provenance data includes using machine learning to summarize event sequences to tell stories of analytic processes. For example, machine learning could be used to curate provenance data to abstract sequences into higher level tasks to describe the analysis. Machine learning could be used on provenance logs to identify canonical analytic workflows and generalize them through fuzzy matching. This could be useful for creating data-driven descriptive theory about real-world analytics work, to compare to idealized models of visual analytics [3].

If models of analytic workflows can be created from event logs, machine learning might then be helpful in mixed-initiative interfaces providing guidance to analysts and users of visualization systems. Such guidance may come in the form of recommended next steps in an analysis, based on learned effective analytic sequences (data support) or on learned usage patterns for specific analytic software (interface support).

Furthermore, interface and analytic process customization may be possible if machine learning on provenance logs can be leveraged to draw conclusions about user characteristics, both long-term traits (e.g., personality factors, locus of control, level of analytic expertise) and

short-term transient traits (level of stress, level of cognitive load). If these user characteristics could be determined to a reasonable level of confidence, a system may make recommendations of analytic next steps, or interface support, which may be specific to the user.

### 4.4.2 Provenance for Machine Learning

The second way that provenance and machine learning can come together in visual analytics is in generating provenance data during the generation of machine learning models, essentially to support model creation, refinement, and deployment. For example, it could be useful to record the steps to develop a model, the parameter settings used both in experiments and in the final model, the features used to train the model, and the steps of model validation.

Furthermore, provenance of machine learning models could be useful in the creation of explainable machine learning visualizations. A classifier, for example, may reveal the features and input data examples which are most influential in a given classification trial. Another example would be a visualization which reveals the sensitivity of the classifier to the specific value of parameters driving the model.

### 4.4.3 Case Study: Log Event Chunking

A fundamental problem underpinning many applications of visual analytics provenance is summarizing the high-volume, low-level event data into a smaller, more cognitively manageable, and semantically meaningful set of chunks, which we will refer to as tasks [1]. Given a sequence of interaction events from a visual analytics system, the problem is to group, segment, or chunk the events into subsets corresponding to higher-level tasks (see Figure 3). Here we focus on the problem of automated or semi-automated chunking and leave the problem of labeling or identifying the higher-level meaning of the resulting chunks to future work.

There are several challenging aspects of the chunking problem:

- The notion of higher-level tasks (or goals or intent) is not precise or may not be known a priori, making top-down aggregation difficult.
- Tasks may be hierarchical, with tasks being themselves part of higher-level tasks or consisting of subtasks, with the steps that correspond to the interaction events in the log being just the lowest level.
- Tasks may be interleaving (e.g., the user starts a task, interrupts it to switch to a different task, then switches back to the original task) or overlapping, so that the boundaries between tasks may be fuzzy, and tasks may consist of non-adjacent events.

#### 4.4.3.1 Possible Approaches

One potential formulation of the problem is to label each event with the chunk that it belongs to. However, this solution would suggest stricter boundaries between chunks and more certainty in the chunk assignments than is present in reality, given the inherent ambiguity in tasks, as described above. A way of remedying this is to augment each event's chunk assignment with a probability that corresponds to the confidence of that assignment. Going one step further, providing the probability, for each event, that the event belongs to each cluster, would provide even more information and would more accurately reflect the characteristic ambiguity described above.

With this formulation we can consider whether to use a supervised, unsupervised, or semi-supervised machine learning approach.

| Event | Group |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |
| 7 | 2 |
| 8 | 2 |
| 9 | 2 |

(a)

| Event | Group | Confidence |
|---|---|---|
| 0 | 0 | 1.00 |
| 1 | 0 | 0.66 |
| 2 | 1 | 0.32 |
| 3 | 1 | 0.58 |
| 4 | 1 | 0.93 |
| 5 | 1 | 0.83 |
| 6 | 1 | 0.67 |
| 7 | 2 | 0.49 |
| 8 | 2 | 0.73 |
| 9 | 2 | 0.76 |

(b)

| Event | Group 0 | Group 1 | Group 2 | ... |
|---|---|---|---|---|
| 0 | 0.38 | 0.00 | 0.00 | ... |
| 1 | 0.56 | 0.00 | 0.00 | ... |
| 2 | 0.06 | 0.02 | 0.01 | ... |
| 3 | 0.00 | 0.18 | 0.03 | ... |
| 4 | 0.00 | 0.43 | 0.01 | ... |
| 5 | 0.00 | 0.23 | 0.00 | ... |
| 6 | 0.00 | 0.67 | 0.16 | ... |
| 7 | 0.00 | 0.14 | 0.22 | ... |
| 8 | 0.00 | 0.01 | 0.34 | ... |
| 9 | 0.00 | 0.00 | 0.23 | ... |

(c)

**Figure 3** Three ways of grouping events. In (a) events are each assigned to a group. In (b), these assignments are augmented with a confidence level, to reflect the inherent ambiguity of the task. In (c), events are given a probability of belonging to each group, allowing for overlapping groups.

- For supervised algorithms, we would require a set of chunk labels. However, there is no natural source of these labels, and manually generating the many labels that would be needed is tedious. Moreover, labelers will be inconsistent with one another for the reasons described above regarding the inherent ambiguity of tasks, making it important to obtain labels from a sufficiently broad sample of labelers.
- For unsupervised algorithms, we dispense with the need for labels, however, the challenge is to develop a meaningful similarity or distance measure such that events in the same chunk register as more similar or closer together and events in different chunks, further away.
- Semi-supervised learning can combine the strengths of both of these approaches by using both a similarity measure and selective use of labels to reduce ambiguity in borderline cases; in this way fewer labels are needed, and weaknesses of the similarity measure can be compensated for. However, semi-supervised learning techniques can be sensitivity to inconsistent labels (e.g. a labeler providing different labels for the same data item during a session), making the outcomes less repeatable or stable. Another challenge would be in designing a label solicitation interface that provided enough context to the labeler.

### 4.4.4 Potential Algorithms

Given a machine learning approach (supervised, unsupervised, or semi-supervised), we suggest a number of possible algorithms that could be suitable to perform the segmentation of interaction logs. This list is not meant to be exhaustive, but to provide the initial seeding to the design and adoption of techniques from other communities.

- Frequent sequential patterns
- Bioinformatics sequence algorithms
- Regex pattern generation
- Program synthesis (e.g. use of domain specific language)

- Fuzzy clustering or classification
- Crowdsourcing
- Bayesian modeling

### 4.4.5 Provenance Data to Machine Learning Features

Determining what features of interaction records and system events may be useful in the chunking requires some feature engineering. Many interactive systems are instrumented for some sort of event logging, but in the following section we enumerate a variety of feature types which could be added to instrumented software to provide a better set of features for chunking. Our list makes use of "basic features" in general interactive systems as well as features that are specific to visualization and visual analytics systems.

#### 4.4.5.1 Basic Features

Interaction records of software, including visualization systems, often record timestamped low-level events such as mouse movements, clicks, keys typed, use of interface functions such as undo, redo, and buttons. Based on these low-level events, additional information can be derived. For example, the velocity of the mouse movement, the fact that a user performed a selection bounding box (through click-and-drag).

These particular features have to be carefully curated so that the logs collect information useful for machine learning on provenance data, but do not contain or reveal personally identifiable information. For example the velocity of typing, or the classes of keys (letters/numbers) or words (stop words/content words) may be recorded without recording the actual content typed into the logs.

#### 4.4.5.2 Visualization Specific Features

Event records specific to visualization applications include, most importantly, those that relate to the data. For example, the use of filter tools to filter a dataset, including the filter parameters, would be a feature to log in the interaction records. In addition, the actual data visible on screen at any moment would be useful to log either in association with all events, or whenever the visible data changes. However, logging all visible data may introduce storage considerations if the dataset is large. Some systems may have succinct ways to describe the dataset (e.g. hashing, lists of constraints and filters, etc.). It may also be possible to capture this information through descriptive features of the data, such as mean, standard deviation, presence of outliers, distribution of node degree (for graphs), etc. Or, a low-resolution screenshot of the visualization state may suffice as a proxy for the list of visible data items. Otherwise, if the entire visible dataset must be recorded, we leave the challenge of addressing the storage problem to future work.

In a coordinated multi-view visualization system, which visualization panel is currently active may be appropriate to place in the log. Along with the active panel, lower-level focus events could be logged, such as mouse hover on visual items (data items and visualization features such as the axes). Depending on the availability of additional interface hardware, such as eye-tracking, dwell time on visual items could also be logged. Analytic actions such as annotation (specific to data items) and note-taking (general about a visualization state) should also be recorded in the log.

#### 4.4.5.3 Novel Features to Log

Moving forward from the more traditional interaction records, machine learning, and specifically chunking, on provenance logs may be more successful if we log new types of features specific to analytics systems. First, we may encode the system state in a feature vector, or potentially reduce the system state and data state to a point [4, 2] which can be compared to other points to determine a distance from previous states.

While recording the data which is on screen and also which has been explicitly of interest through hover or selection events can be useful, it may be possible to discover features in the logs which are more closely tied to the task of analytics. For example, statistical relations of focused (hovered) data to other data could indicate high-level interest patterns. Are outliers being hovered? Are items of interest in a cluster? Another feature could be to track the number of recently visited items affected by an operation. Does the filter remove recently hovered items? Then the filter action is probably part of a sequence. Image-based measures targeted toward visualization could also be informative for discovering important moments in analysis, such as the amount of change in the displayed image (e.g. image-based or model of perceived change). Similarly, back-end logs such as data load operations could indicate major changes in direction in an analysis process.

### 4.4.6 Summary

In this report, we describe the mutual benefit between provenance and machine learning, and focus on a particular problem that is of relevance for both – chunking of event data. We discuss possible machine learning approaches and list some algorithms that have potential in addressing the problem. More concretely, we brainstorm novel features, besides standard keyboard and mouse events, that could be considered in future chunking solutions.

#### References

1 Sara Alspaugh. *Understanding Data Analysis Activity via Log Analysis*. PhD thesis, EECS Department, University of California, Berkeley, Aug 2017.

2 Eli T Brown, Sriram Yarlagadda, Kristin A Cook, Remco Chang, and Alex Endert. Modelspace: Visualizing the trails of data models in visual analytics systems. In *Workshop on Machine Learning from User Interaction for Visualization and Analytics*, pages 1–11, 2018.

3 Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. pages 2–4, 2005.

4 S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk. Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):1–10, Jan 2016.

5 K. Xu, S. Attfield, T. J. Jankun-Kelly, A. Wheat, P. H. Nguyen, and N. Selvaraj. Analytic provenance for sensemaking: A research agenda. *IEEE Computer Graphics and Applications*, 35(3):56–64, May 2015.

## 4.5 Storytelling & User Intent

*Robert Kosara (Tableau Software – Seattle, US), Ali Sarvghad (University of Massachusetts – Amherst, US), William Wong (Middlesex University – London, GB), and Michelle X. Zhou (Juji Inc. – Saratoga, US)*



While visualization research tends to assume that the exploration and analysis of the data come before the presentation is designed, the analysis is often motivated and guided by an assumed story, hunch, or hypothesis. Analysts might go back and forth between story construction and analysis many times along the way, discovering new questions to ask as they create the story, finding supporting or contradicting evidence in their analysis, etc.

### 4.5.1 Levels of Intent

A central challenge in understanding user intent and incorporating it into the storytelling process is the mapping between low-level interaction events and higher-level goals of an analysis. We described four levels of abstraction in user intent: Operation: The lowest-level of interaction event (mouse click, selection event) in a visual interface Task: The activity of which an operation is a part (e.g., a selection event coupled with drag-and-drop may be part of a "categorize" task) Goal: The reason a task is performed Intent: The highest-level abstraction for the purpose behind an analysis (or analysis subsequence)

Operation and Task together comprise what an analyst does. Goal and Intent together comprise why. A complete taxonomy of intent levels should borrow from the human factors literature on hierarchical task analysis [4].

Stories can communicate the intent behind an analysis (by representing the choices and interpretations an analyst made) to support interpretation by the analysis consumer. Stories can also drive intent in that they provide a template for how a data-driven argument was constructed and allow an analysis producer to reuse a story structure with new data.

### 4.5.2 From Provenance to Story

Provenance can support the construction of the presentation by surfacing states in the analysis that are likely of interest. Heuristics for selection include node centrality of the state within the provenance graph, repetition of states (a state that was visited more often is more likely to be of importance), amount of change from the previous state, explicit user tags, etc.

Once potentially useful states have been identified, the user can select which ones to include in the story and insert them into a story structure. At this point, the type of story or argument can be used to select pre-defined story templates according to a number of classic rhetorical structures, such as persuasion, argument, analysis, or exposition. More corresponding structures and prototypes need to be identified, but recent work has shown at least one distinct and reusable pattern for data-based arguments in news graphics [2].

Story structure may be specified ahead of time via a template but may also emerge over time as an analyst explores data. Emergent stories themselves have provenance, which reflects the evolution of an analyst's argument or explanation over time.

Grice's *conversational maxims* [3] describe other desirable properties of the structure and content of stories. The maxims of quantity, quality, relation, and manner provide guidelines for coherent conversation (such as the asynchronous communication between an analysis producer and consumer) that, when violated, introduce ambiguity or other rhetorical flaws. Storytelling interfaces may enforce these maxims as an aid to constructing strong data-based arguments.

### 4.5.3 Structuring Arguments and the Circle

Toulmin provides an explanation of how arguments may be set out so that claims that are made have a basis, and that evidence can be provided to support those claims. Toulmin's argumentation model [5] is a useful way of thinking about a generic structure that can be adapted by a variety of different approaches to representing that structure for purposes for communicating and presenting information. The CFO model [2] is one such useful approach to organizing materials for presentation. The Toulmin model is also useful as it includes other factors that could be considered when we collect data and results of analysis to communicate our findings, e.g. the concept of warrants is the authority on which claims are made. These can take the form of higher order assumptions upon which our explanation is based.

One approach taken based on some of these ideas has been reported in Groenwald, et al. [1]. The approach to storytelling in this example is to construct the story made up of data elements and results of analyses into unique sequences that help to explain what the analyst has observed in the data. The ideas generated by the initial sequence of data can be used to tell communicate a story – an explanatory narrative. This initial understanding provides the basis for formulating an early stage tentative hypothesis – a hunch – that can guide further inquiry, and more data collected to prove or disprove the hypothesis.

This process informs the communicator/analyst, develops new or elaborates his understanding, which enables him to seek further data or analyses, question its findings, and to even reframe his conceptualisation of the problem and the way he intends to communicate the message [6].

### 4.5.4 Next Steps

We plan to further investigate how provenance data and visualisations might be used to support the process of communicating the outcomes from analyses:

- Use provenance trails to highlight nodes of interest
- Select nodes for story
- Assemble the nodes into rhetoric structure
- Add a narrative to create an explanation

Logging user intent provides a promising starting point for a wealth of research into building more effective presentations and stories.

#### References

1 Celeste Groenewald, Simon Attfield, Peter Passmore, B. L. William Wong, Nadeem Qazi, and Neesha Kodagoda. A descriptive, practical, hybrid argumentation model to assist with the formulation of defensible assessments in uncertain sense-making environments: an initial evaluation. *Cognition, Technology & Work*, 20(4):529–542, Nov 2018.

2 Robert Kosara. An argument structure for data stories. In *Short Paper Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)*, 2017.

3 Stephen Neale. Paul grice and the philosophy of language. *Linguistics and philosophy*, 15(5):509–559, 1992.

4 Neville A Stanton. Hierarchical task analysis: Developments, applications, and extensions. *Applied ergonomics*, 37(1):55–79, 2006.

5 Stephen Toulmin. The uses of argument. 2003.

6 B. L. W. Wong, P. Seidler, N. Kodagoda, and C. Rooney. *Supporting variability in criminal intelligence analysis: From expert intuition to critical and rigorous analysis*, pages 1–11. Springer International Publishing AG.

## 4.6 User Modeling & Intent

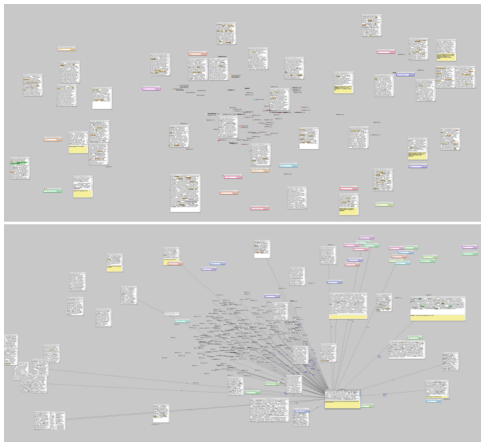
*John Wenskovitch (Virginia Polytechnic Institute – Blacksburg, US), Remco Chang (Tufts University – Medford, US), Christopher Collins (UOIT – Oshawa, CA), Michelle Dowling (Virginia Polytechnic Institute – Blacksburg, US), Alex Endert (Georgia Institute of Technology – Atlanta, US), Phong H. Nguyen (City – University of London, GB), Kai Xu (Middlesex University – London, GB), and Michelle X. Zhou (Juji Inc. – Saratoga, US)*



### 4.6.1 Motivation

Inferring user intent from interactions is an active problem in visual analytics. Current solutions to this problem in the realm of interactive projections make use of online learning, inferring the intent of individual interactions to incrementally build a user interest model.

For an example, consider a visual text analytics system. In this system, a collection of documents are laid out spatially using a distance similarity metaphor – similar documents are positioned close together, while dissimilar documents are positioned further apart. The system responds to user interactions by updating an interest model (if a user places two documents close together, the system identifies what makes these documents similar and thereby learns what factors the user is interested in), and may forage for new relevant information based upon what the system learned. An example of this foraging can be found in the accompanying figure.



However, this system may need to respond differently depending on the skill level and traits of the users. For example:

- Users could be experienced system users (e.g., power users are ready to be overwhelmed with new information that they can easily process) or they could be novices (e.g., let's not give them too much information at once until they are comfortable with the system).
- Users could be intellectually curious (e.g., open to exploring a number of conflicting hypotheses) or they could want hand-holding (e.g., explore fewer possibilities).

An open question is how the system can obtain or learn this information about its users: whether it comes from monitoring and interpreting user behavior, or is detected from a survey before using the system, or another means entirely.
Our group's discussion on Tuesday sought to explore some of the options in this space.

### 4.6.2 Modeling Characteristics vs. Modeling Intent

To begin, we developed two separate lists: (1) a list of user characteristics that could affect the system preferences of a user, and (2) a list of goals that a system designer may want to include in their software. These are not necessarily complete lists, but they gave us a starting place to work from. These lists follow:

**User Characteristics**

1. Independence
2. Expertise
3. Level of stress
4. Adaptability
5. Need for control
6. Intellectual curiosity
7. Dark Triad
8. Locus of control
9. Tolerance of uncertainty

**System/Designer/User Goals**

1. Avoid biases
2. Information and analysis coverage
3. Efficiency
4. Detect if user needs help/reassurance
5. Understand how users use a system (post-hoc)
6. Storytelling, report generation
7. Delivering different data (in different ways) for different types of users
8. User happiness ("the system gets me")
9. Increase user awareness of the implication of their behaviors/analysis (e.g., ethics, biases, assumptions)

It is worth noting that the user characteristics listed to the left have a variety of temporal spans. For example, a user's level of stress could fluctuate during the time spent using the system, whereas expertise is more constant but dependent on the subject area of analysis, and intellectual curiosity is a still more constant behavior.

The mapping between these characteristics and goals is also a bit nebulous. For example, the (3) level of stress user characteristic could map to a number of goals. Most obvious is (4) detecting if a user needs help of reassurance; if the stress is due to issues with understanding the system or the data, the system may wish to reduce the rate of information flow to a simpler level, or perhaps may pop up some tips for how to continue with the analysis. Reducing the rate of information flow therefore also effects (2) information and analysis coverage, while redirecting the analysis path effects (1) avoid biases and (7) delivering data for different types of users (as a stressed user is different from a non-stressed user). Analysis of how a user responds to this reduction in information flow further triggers goal (5) and possibly also (8).

These interventions in response to user behaviors could either be handled by the front-end or the back-end of a system. There is, of course, a tradeoff inherent in these intervention options:

- **Front-end:** There could be different levels of visibility for the intervention (a spectrum from subtle hints to locking out some system functionality), but each necessitates an interruption to the user's workflow.
- **Back-end:** Low risk and no obvious interruption, but failure of a predicted intervention is not visible to the user.

We also discussed how different user models can be classified and characterized. We decided that there are three types of predictive user models:

1. **Understand intent:** These user models determine the interests of a user based on their interactions. This could serve to adapt the behavior of the system as suggested above (e.g., change the rate or variety of information flow in response to learned characteristics).
2. **Predict future user actions:** These user models predict future interactions for users during their analysis process (e.g., to increase system response time by preprocessing future analysis or to suggest analysis routes to a user).
3. **Classify user characteristics:** These user models can assist with post-hoc analysis of system behavior for future versions (e.g., learn what types of users most often use the system so that menus and toolbars can be organized for ease of access to common features), and can also assist the other two user model types.

### 4.6.3 Using Semantic Interaction to Model Characteristics

Following our discussion on possible characteristics and system goals that might be included in future systems, we turned our attention to how a system could obtain this information. The example of giving a user a pre-survey to understand the user is useful but also trivial and potentially misleading (e.g., users don't want to admit their lack of expertise or current stress level). Our discussion instead focused on if the Semantic Interaction paradigm, intended to infer user intent, could be adapted to learn user characteristics. For example:

| **Semantic Interaction** | **Modeling Characteristics** |
|---|---|
| 1. Capture an interaction | 1. Capture an interaction |
| 2. Interpret intent | 2. Interpret *characteristics* |
| 3. Update the model | 3. Update the model |

Step 2 on the right is the challenge that needs to be addressed in order to build user models that learn these user characteristics.

### 4.6.4 Mapping Interactions to Intent and Characteristics

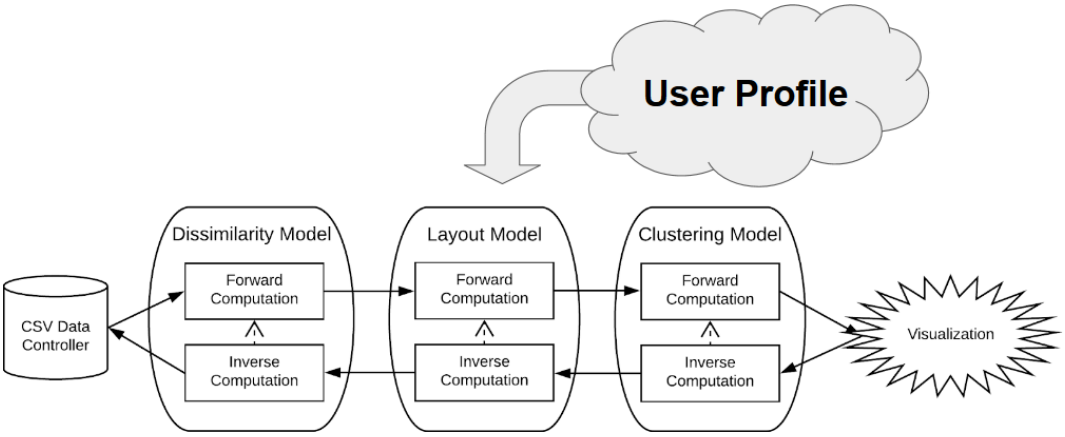
A necessary step in both semantic interaction and in modeling these user characteristics is the inference phase. Without clear-cut rules, inferring the intent of a user based only upon interactions is a clear challenge. In discussing this problem, we structured our discussion by cardinality of both interaction and intent:

- **One interaction implies one intent:** This is the trivial case, and one example would be the direct manipulation of a control widget (click the button to submit the form).
- **Many interactions imply one intent:** This case can be thought of as flexible UI design. For example, there are many different interactions and interaction sequences that can be used to bold text in Microsoft Word. Though the intent is the same, a separate set of interactions may be supported in OpenOffice to achieve the same goal.
- **One interaction implies many intents:** This case is an underspecified interaction: one interaction from a user could be inferred in many different ways. We discussed five different ways to disambiguate this uncertainty in interaction:
  1. **Ask the user to disambiguate:** A simple case of popping up a "What did you intend this interaction to do?" message.
  2. **User provenance to infer the user's intent:** Given a past sequence of interactions, we could attempt to guess at the most likely intent of an interaction.
  3. **Get more examples from the user:** An ambiguous interaction may not alter the underlying model until it is performed many times, or until the user broadens the scope of the interaction (e.g., flash-fill in Excel).
  4. **User data to infer:** Similar to (2), but using the dataset under consideration rather than the past interactions of the user.
  5. **Use user characteristics to infer:** Similar to (2) and (4), but using what the system has learned about the behavior and characteristics of the user to disambiguate.
- **Many interactions imply many intents:** This is the most interesting case, because the obvious (to our discussion) interpretation is flexible interaction design. The user could perform any gesture or interaction, and the system could use a set of meta-rules (or user behavior, or provenance, or any of the above) to infer the user's intent for the interaction.

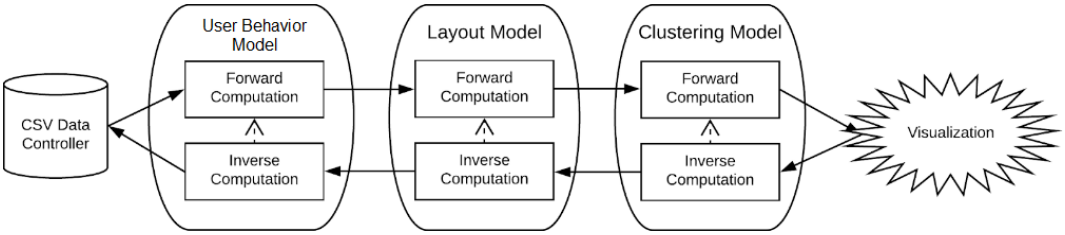
Though we did not discuss this in detail, a similar cardinality breakdown could be used to map interaction to learned user behaviors. For example, shaking the mouse in frustration could easily map to an increase in the inferred frustration level of a user.

### 4.6.5 How This Could Work / Building a System

After inferring user characteristics, we want to know how best to use these to influence and adapt existing and new systems. Our initial solution was to keep user characteristics as a separate set of parameters that can influence each of the individual components of a system. For example:




However, it is also possible to build this user model into the system as a component that is only processed when it is necessary to reference.




### 4.6.6 Other Issues Discussed

The discussion provided above is a shortest path through relevant topics that we discussed, but a number of other subjects were briefly discussed in side conversations and parallel threads. A quick summary of these discussions are included here:

- What signals can we and should we collect? There are a lot of signals, and also a lot of garbage. How can we tell the signal from the noise?
- What role does training have on user intent expression?
- Data scarcity is a problem. How many user logs do you have? You don't have tens of thousands of users to evaluate. Can we built a model on a single user while they're learning about a system for the first time?
- What is the purpose of this modeling in general? Do we want to build a single-use model, or a persistent model? This is a key difference between post-hoc and real-time learning.
- How do we avoid being Clippy? Can we always get it right, and without being disruptive?

- Struggle to balance instruction vs. freedom to explore your interfaces.
- Capturing low-level parameters is much more tractable than predicting the user's intent.
- Can we specify how many levels of intent we want? Or just low-level and high-level?

## Participants


- Sara Alspaugh
Splunk Inc. – San Francisco, US
- Daniel Archambault
Swansea University, GB
- Simon Attfield
Middlesex University – London, GB
- Leilani Battle
University of Maryland – College Park, US
- Christian Bors
Technische Universität Wien, AT
- Remco Chang
Tufts University – Medford, US
- Christopher Collins
UOIT – Oshawa, CA
- Michelle Dowling
Virginia Polytechnic Institute – Blacksburg, US
- Alex Endert
Georgia Institute of Technology – Atlanta, US
- Jean-Daniel Fekete
INRIA Saclay – Orsay, FR
- Melanie Herschel
Universität Stuttgart, DE
- T. J. Jankun-Kelly
Mississippi State University, US
- Andreas Kerren
Linnaeus University – Växjö, SE
- Steffen Koch
Universität Stuttgart, DE
- Robert Kosara
Tableau Software – Seattle, US
- Olga A. Kulyk
DEMCON – Enschede, NL
- Robert S. Laramee
Swansea University, GB
- Sérgio Lifschitz
PUC – Rio de Janeiro, BR
- Aran Lunzer
OS Vision – Los Angeles, US
- Phong H. Nguyen
City – University of London, GB
- William Pike
Pacific Northwest National Lab. – Richland, US
- Ali Sarvghad
University of Massachusetts – Amherst, US
- Claudio T. Silva
New York University, US
- Holger Stitz
Johannes Kepler Universität Linz, AT
- Melanie Tory
Tableau Software – Palo Alto, US
- John Wenskovitch
Virginia Polytechnic Institute – Blacksburg, US
- William Wong
Middlesex University – London, GB
- Kai Xu
Middlesex University – London, GB
- Michelle X. Zhou
Juji Inc. – Saratoga, US


Report from Dagstuhl Seminar 18471

# Next Generation Domain Specific Conceptual Modeling: Principles and Methods

**Edited by**

**Heinrich C. Mayr[1], Sudha Ram[2], Wolfgang Reisig[3], and Markus Stumptner[4]**

1 **Alpen-Adria-Universität Klagenfurt, AT,** heinrich.mayr@aau.at
2 **University of Arizona – Tucson, US,** ram@eller.arizona.edu
3 **HU Berlin, DE,** reisig@informatik.hu-berlin.de
4 **University of South Australia – Adelaide, AU,** markus.stumptner@unisa.edu.au

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18471 "Next Generation Domain Specific Conceptual Modeling: Principles and Methods". It summarizes the resultsof the seminar and shows in which direction (Domain Specific) Conceptual Modeling shoulddevelop in the opinion of the participants. In addition, the report contains abstracts of thenumerous talks presented during the seminar as well as a summary of the discussions held inworking groups during the seminar. In particular, some open questions will be touched upon,which will be dealt with before a follow-up seminar.

**Seminar** November 18–23, 2018 – http://www.dagstuhl.de/18471
**2012 ACM Subject Classification** Applied computing → Enterprise computing, Computing methodologies → Knowledge representation and reasoning, Computing methodologies → Modeling and simulation, Software and its engineering → Software system models, Software and its engineering → System description languages, Theory of computation → Data modeling
**Keywords and phrases** DSML, Meta Model, Modeling, Modeling Method, Semantics
**Digital Object Identifier** 10.4230/DagRep.8.11.63
**Edited in cooperation with** Pia Wilsdorf, Universität Rostock, Germany

## 1 Executive Summary

*Heinrich C. Mayr (Alpen-Adria-Universität Klagenfurt, AT)*


**Joint work of** The seminar participants and organizers

Models are the basic human tools for managing complexity and understanding and therefore play a key role in all scientific and engineering disciplines as well as in everyday life. Many modeling paradigms have evolved over time into a wide variety of modeling languages, methods and tools that have come and gone. This is particularly true for Informatics, which is a modeling discipline in itself.

Since the 1970s, special attention has been paid to conceptual modeling. This approach essentially uses a formal language whose concepts are linked to a semantic interpretation (e.g. by the grounding in an ontology) and a more or less transparent graphic or textual representation (which supports efficient linguistic perception). Normally, such a language


is embedded in a model/meta model hierarchy. The dimensions of conceptual modeling languages are structure, dynamics (behavior) and functionality.

Despite all efforts, however, there is still no comprehensive and consistent use of conceptual modeling in practice. Often conceptual models are only used as prescriptive documents, which – e.g. in the area of software development or business process management – are rarely synchronized with the developed artifact, so that reality and model diverge step by step. This observation motivated us to promote and conduct this seminar by focusing on domain-specific conceptual modeling, as this promises a methodology that is more tailored to the needs of each user group.

In view of the highly committed discussions during the seminar, the intensive discussions in the working groups and the very positive results of the participant survey, we can say without exaggeration that the seminar was a complete success. Almost all participants wished for a continuation, which we will probably apply for in 2021, when the already decided projects (cooperation and publications of subgroups) are on their way.

Since, with one exception, every participant wanted to present their ideas in a talk, the programme was tightly packed: 36 talks and 2 full evening sessions in working groups, the results of which were reported on the next morning, made the seminar week a very intensive but also highly inspiring experience.

First results are already tangible:

- The workshop "Conceptual Modeling for Multi-Agent Problem Solving" at the IJCAJ 2019 in Macao: The idea was born during the seminar and implemented afterwards: http://austria.omilab.org/psm/content/cmmaps19/info.
- A working group is currently writing a summary paper on the results of the working groups on which agreement was reached in the plenary discussions.
- Questions that were discussed during the seminar will be included in contributions to the Summer School "Next Generation Enterprise Modelling in the Digital Transformation" in Vienna (July 15-26, Vienna).
- The seminar organizers are currently writing a somewhat more popular scientific column to be submitted to CACM.

A number of open questions and "grand challenges" that also could be topics of future relevant conferences have been identified, among others:

- Business Transformations in the age of digitalization as "Models are driving the Digital Transformation"
- Social Aspects of Conceptual Modeling
- Explanatory Models for Neural Networks and Big Data
- Conceptual Modeling for validation purposes in simulation
- Modeling of Ultra Large Scale Architectures
- Privacy Modeling
- Modeling of Behavior Goals for Assistive Systems and Emotions
- Better integration into teaching at universities of applied sciences and universities
- Tools and Technical Infrastructures for Conceptual Modeling, in particular for "multi-metamodeling frameworks"
- Involvement of researchers and practitioners from other fields: "go beyond the obvious".

The biggest challenge for a follow-up seminar will be to encourage more practitioners to participate. For this purpose, we will propose to dedicate two consecutive seminar days to this and the discussion with them, as practitioners usually cannot spend more time.

## 2 Table of Contents

# 3 Overview of Talks

## 3.1 Multi-Level (Domain-Specific) Conceptual Modeling

*João Paulo Almeida (Federal University of Espírito Santo – Vitória, BR)*



Conceptual models are often built with techniques which propose a strict stratification of entities into two classification levels: a level of types (or classes) and a level of instances. Despite that, there are several situations in which domains of inquiry transcend the conventional two-level stratification and domain experts use types of types (or categories of categories) to articulate their conceptualizations. For instance, in a project we are currently involved in–concerning integration of water quality data in the Rio Doce river basin [1]–the ontology-based conceptual models we are defining must cover both particular water quality measurements (observations set in a particular time and location) as well as the types of measurement they instantiate ("water sampling", "soil sampling", "specimen sighting", "specimen collection"); types of aquatic animals ("native species", "invasive species", the various types of fish according to biological taxonomy and systematics: "pimelodid catfish", "red piranha") as well as specific specimens (e.g., a specific catfish collected for analysis).

In these settings, types are instances of other types and multiple levels of classification can be identified (individuals, classes, metaclasses, metametaclasses, and so on), characterizing what is now called "multi-level modeling" [2].

In my talk, I have discussed how multi-level conceptual models are relevant not only in the conceptual modeling of specific domains (as illustrated earlier), but also in the definition of the real-world semantics of (domain-specific) modeling languages. In this process, it is key to identify that we are addressing two tasks during modeling language engineering: the design of a language's abstract syntax (often approached by defining a metamodel and associated syntactic constraints) and the definition of the language's semantics in terms of a reference ontology [3].

**Acknowledgments.** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

### References

1 Patrícia M. Campos, Cássio C. Reginato, João Paulo A. Almeida et al., *Building an Ontology-Based Infrastructure to Support Environmental Quality Research: First Steps*, Proc. XI Seminar on Ontology Research in Brazil, São Paulo, Brazil, 2018, pp. 227–232.

2 João Paulo A. Almeida, Ulrich Frank and Thomas Kühne, *Multi-Level Modelling (Dagstuhl Seminar 17492)*. Dagstuhl Reports, vol. 7, 2018, pp. 18–49.

3 Victorio A. Carvalho, João Paulo A. Almeida, G. Guizzardi, *Using Reference Domain Ontologies to Define the Real-World Semantics of Domain-Specific Languages*, Proc. 26th International CAiSE Conference, 2014, pp. 488–502.

## 3.2 On the Quality of Requirements Goal Models

*João Araújo (New University of Lisbon, PT)*


**Joint work of** Catarina Gralha, João Araújo, Miguel Goulão, Mafalda Santos, Ana Moreira
**Main reference** Catarina Gralha, João Araújo, Miguel Goulão: "Metrics for measuring complexity and completeness for social goal models", Inf. Syst., Vol. 53, pp. 346–362, 2015.
**URL** https://doi.org/10.1016/j.is.2015.03.006

Requirements models have been developed for the requirements engineers and stakeholders work, providing abstraction mechanisms to, for example, facilitate the communication among them by providing better structuring of requirements, thus helping with their analysis. Nevertheless, the extent to which requirements modelling languages are adequate for communication purposes has been somewhat limited. Several quality aspects have contributed to that, ranging from lack of abstraction mechanisms to address model's complexity, to the impact of layout of models or the actual notation adopted. For example, in large-scale systems, building requirements models may end in complex and/or incomplete models, which are harder to understand and modify, leading to an increase in costs of product development and evolution. Consequently, for large-scale systems, the effective management of complexity and completeness of requirements models is vital. Moreover, it is undeniable that the communication potential of requirements modeling languages is not entirely explored, as their cognitive effectiveness is often not boosted. For example, choosing an ad-equate layout for requirements models may be a relevant issue, as a bad layout may compromise the adequacy of the models. Also, although visual notations are often adopted (as they are perceived as more effective for conveying information to nontechnical stakeholders than text), their careful design is often not considered. Not taking all this into account, in the long run, may result in poorly understood requirements, leading to problems in artifacts produced in later stages of software development. So, in this talk, I will discuss in detail these issues based on the results of experiments where metrics were collected to evaluate and discuss some quality aspects of requirements models, in particular requirements goal models (increasingly popular in the requirements community), such as complexity, completeness, understandability and semantic transparency.

## 3.3 Conceptual Modeling Issues in Knowledge-intensive Processes

*Fernanda Baião (PUC – Rio de Janeiro, BR)*


**Main reference** Juliana Baptista dos Santos França, Joanne Manhães Netto, Juliana do E. Santo Carvalho, Flávia Maria Santoro, Fernanda Araujo Baião, Mariano Gomes Pimentel: "KIPO: the knowledge-intensive process ontology", Software and System Modeling, Vol. 14(3), pp. 1127–1157, 2015.
**URL** https://doi.org/10.1007/s10270-014-0397-1
**Main reference** Tiago Prince Sales, Fernanda Araujo Baião, Giancarlo Guizzardi, João Paulo A. Almeida, Nicola Guarino, John Mylopoulos: "The Common Ontology of Value and Risk", in Proc. of the Conceptual Modeling – 37th International Conference, ER 2018, Xi'an, China, October 22-25, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 11157, pp. 121–135, Springer, 2018.
**URL** https://doi.org/10.1007/978-3-030-00847-5_11
**Main reference** Bedilia Estrada-Torres, Pedro Henrique Piccoli Richetti, Adela del-Río-Ortega, Fernanda Araujo Baião, Manuel Resinas, Flávia Maria Santoro, Antonio Ruiz-Cortés: "Measuring Performance in Knowledge-intensive Processes", ACM Trans. Internet Techn., Vol. 19(1), pp. 15:1–15:26, 2019.
**URL** https://dl.acm.org/citation.cfm?id=3289180

Knowledge-intensive Processes, simply put, are a composition of prospective activities (events) whose execution contributes to fulfilling a goal and whose control-flow, at the instance level, typically presents a high degree of variability. KiPs are acknowledged as the most valuable

assets in current organizations; nevertheless, there are several elements which, apparently, impact their behaviour in an unpredictable way, posing risks that are difficult to be managed. Therefore, KiPs pose several challenges with regard to an ontology-based definition of what they essentially "are", a corresponding metamodel that is able to constrain the set of possible KiP models that may be generated, adequate modeling languages to represent them in such a way to provide adequate understanding, assessment and management by process stakeholders, and a technological infrastructure that is able to keep track of observed instances from the real-world.

In this talk, I point to a set of initiatives that were/are being conducted in our research group to address each of this challenges. These initiatives are organized in a framework called KiPaIS (Knowledge-intensive Process-aware Information System), which comprises: (i) CognitiveKiP, a cognitive-based ontology for KiPs; (ii) KiPO, a metamodel for KiP modeling that applies Multi-Level modelling and combines declarative and imperative modelling approaches; (ii) KiPN, a graphical modeling language for the domain of KiPs, and (iv) KiPOwl, a codification of KiPO in OWL, stored in a NoSQL database, to enable instantiation of KiP instances from several sources, such as documents, declarative modeling tools, event and message logs from transactional systems.

## 3.4 Value-driven Approach for BI Application Design

*Ladjel Bellatreche (ENSMA – Chasseneuil, FR)*



In a very short time, the data warehouse (DW) technology has gone through all the phases of a technological product's life: the introduction on the market, growth, maturity, and decline. Maturity means there is a clearly identified design life cycle plus a race and competition between companies to increase their decision-making power. The decline was signaled by the appearance of Big Data. It is therefore essential to find other challenges that will contribute to the revival of DW while taking advantage of the V's of Big Data. The arrival of Linked Open Data (LOD) era is an excellent opportunity for both the DW academia and industry communities. LOD may bring an additional Value that the sources feeding a DW typically do not usually succeed to yield. Offering the added value of a DW is related to a high Variety of sources. In this talk, first, we show the role of conceptualization to deal with the variety of internal and external sources and study its impact on the ETL phase to ease the value capturing. Secondly, three scenarios related to added value for integrating LOD in the DW are given. Finally, experiments are conducted to show the effectiveness of our approach

## 3.5 Specification Techniques for Conceptual Modeling Methods

*Dominik Bork (Universität Wien, AT)*


**Joint work of** Dominik Bork, Dimitris Karagiannis, Benedikt Pittl
**Main reference** Dominik Bork, Dimitris Karagiannis, Benedikt Pittl: "How are Metamodels Specified in Practice? Empirical Insights and Recommendations", in Proc. of the 24th Americas Conference on Information Systems, AMCIS 2018, New Orleans, LA, USA, August 16-18, 2018, Association for Information Systems, 2018.
**URL** https://aisel.aisnet.org/amcis2018/AnalysisDesign/Presentations/1

Conceptual modeling languages such as BPMN and UML are widely used in industry and academia. Such modeling languages are usually introduced in overarching specification documents maintained by standardization institutions. Being the primary – often even the single – source of information, such specifications are vital for modelers, researchers, and tool vendors. However, how to derive a coherent and comprehensive specification was never systematically analyzed. This presentation reports on the analysis of 11 current modeling language specifications with a focus on how their abstract and concrete syntax are specified. Identified specification techniques are discussed and their sample usage is illustrated. Thereby, individual strengths and weaknesses of each technique are discussed. The contribution of this presentation is a foundation for increasing the consistency and expressive power of modeling language specifications, ultimately leading to an improved understanding and better utilization of those languages.

## 3.6 Conceptual Modeling of Prosopographical Databases

*Isabelle Comyn-Wattiau (ESSEC Business School – Cergy Pontoise, FR)*


**Main reference** Jacky Akoka, Isabelle Comyn-Wattiau, Cedric du Mouza, Stéphane Lamassé: "Modeling Historical Social Networks Databases", in Proc. of the 52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, USA, January 8-11, 2019, pp. 151–161, 2019.
**URL** http://hdl.handle.net/10125/59714

Prosopographical researchers share many concepts: persons, sources of information, location-s/place, time, uncertainty. There is a need to build a common conceptual model putting together these concepts. On such a basis, we could develop more powerful and evolutive databases, available for all historians. In terms of conceptual modeling, there are at least three main challenges: 1) genericity, 2) modeling uncertainty, 3) granularity of factoids/events. I first define prosopography and its differences with connected fields such as onomastics or genealogy. Then I explain why prosopographical researchers need databases and why current databases do not meet all their requirements. In particular, I claim that, except the factoid model, there is no past effort of conceptual modeling for prosopography.

## 3.7 Attribute based communication for Collective Adaptive Systems

*Rocco De Nicola (IMT – Lucca, IT)*


**Joint work of** Rocco De Nicola, Michele Loreti, Yehia Abd Alrahman

I presented the approach I have been following in the past twenty years for developing applications following specific programming paradigms such as: Network Aware Programming, Service Oriented Computing, Autonomic Computing and Collective Adaptive Systems that have brought us to introduce the languages KLAIM, COWS and SCC, SCEL and more recently AbC. I explained what I meant by domain specific formalisms and outlined our approach based on three basic steps (i) introducing a specification language equipped with a formal semantics, (ii) implementing supporting a programming frameworks with its associated runtime environment, (iii) providing verification techniques and tools. After this I concentrated on the Autonomic Computing and Collective Adaptive Systems paradigm and on some of their key notions and advocated the use of a novel communication paradigm that is based on selecting communication partners according to their run time properties expressed as attributes.

## 3.8 Experience in Stochastic Model-Based Dependability Analysis: Modeling and Analysis of Cyber-Physical Systems

*Felicita Di Giandomenico (CNR – Pisa, IT)*



The talk addresses stochastic model-based analysis of critical CPS from the point of view of dependability and energy saving perspective. The goal of the analysis is mainly to assess the impact of faults/attacks (and their propagation) on the ability of the system to provide correct service. A general overview of the conceptual model is first presented, based on the concepts of generality, modularity and compositionality. Then, two application domains are considered: Smart Grids and Railway transportation. Challenges, modeling approaches, property of interests as well as examples of analysis results are briefly discussed. Some general observations on the role of the analysis purpose (e.g., the impact of failures on resilience/QoS related indicators) and the application domain of the system under analysis (e.g., the Smart Grids) in guiding choices leading to conceptual models and model implementations, as well as directions for further research investigations, are drawn at the end.

Dependability of critical infrastructures, such as electric power systems and transportation systems, is paramount, since they provide services our everyday life strongly depends on. Stochastic model-based analysis is a popular approach to assess dependability properties, especially at early stages of system development. However, these infrastructures are Cyber Physical Systems, characterised by a variety of challenging aspects from the modelling point of view, such as: continuos and discrete state variables, failure propagation through interdependencies, heterogeneity and dynamicity of components structure and behaviour, topology-dependent criticality, large size of interconnected components. Based on the principles of generality, modularity and compositionality, a conceptual model can be built, guided by the purpose of the analysis (e.g., the impact of failures on resilience/QoS related

indicators) and by the specific application domain of the system under analysis (e.g., the Smart Grids). However, building such a model is still an art and strongly dependent on the skill and experience of the modeler. A sound approach to assist the model developer in carrying on her/his task in a more rigorous way would be highly desirable. This is identified as a research direction where further investigations are still needed.

### References

1 S. Chiaradonna, F. Di Giandomenico, P. Lollini, Definition, Implementation and Application of a Model-Based Framework for Analyzing Interdependencies in Electric Power Systems, International Journal on Critical Infrastructure Protection, N. 4 (2011), pp. 24–40.
2 Silvano Chiaradonna, Felicita Di Giandomenico, Giulio Masetti A stochastic modelling framework to analyze smart grids control strategies. In: SEGE 2016 – 4th IEEE International Conference on Smart Energy Grid Engineering (Oshawa, Canada, 21-24 August 2016). Proceedings, pp. 123–130. IEEE, 2016.
3 Giulio Masetti, Silvano Chiaradonna, Felicita Di Giandomenico: A Stochastic Modeling Approach for an Efficient Dependability Evaluation of Large Systems with Non-anonymous Interconnected Components. Proceedings 28th IEEE International Symposium on Software Reliability Engineering ISSRE 2017, pp: 46–55
4 Davide Basile, Silvano Chiaradonna, Felicita Di Giandomenico, Stefania Gnesi A stochastic model-based approach to analyze reliable energy-saving rail road switch heating systems. In: Journal of Rail Transport Planning and Management, vol. 6 (2) pp. 163–181. Elsevier, 2016.

## 3.9 The Role of Visualization in Conceptual Modeling

*Hans-Georg Fill (University of Fribourg, CH)*


**Main reference** Hans-Georg Fill: "Visualisation for Semantic Information Systems", Gabler, 2009.
**URL** https://doi.org/10.1007/978-3-8349-9514-8

In many conceptual modeling approaches, the use of visualization techniques is inherent. The used graphical notations are often intuitive and easy to understand, despite the sometimes formal foundation of the modeling languages they are attached to. Visualization thus contributes to the communication of model information and the processing of complex information by humans. Despite existing guidelines that have been proposed for designing graphical notations of modeling languages, this task still requires considerable experience and a good understanding of graphical design and its technical implementation, especially in the case of dynamic notations. The challenge thus persists to provide adequate guidance on designing and choosing good visual representations for models and for simplifying their implementation on metamodeling platforms. Furthermore, when such visualizations are coupled with data-based approaches as found in the area of information visualization, with virtual or augmented reality environments or device-less interaction, a lot of technical know how is required. It should therefore be researched in the future for example, how existing modeling approaches can be transitioned to VR and AR environments and how interaction can take place in such settings.

### References

1 Sandkuhl, Kurt, Fill, Hans-Georg, Hoppenbrouwers, Stijn, Krogstie, John, Leue, Andreas, Matthes, Florian, Opdahl, Andreas, Schwabe, Gerhard, Uludag, Omer, Winter, Robert

(2018): From Expert Discipline to Common Practice: A Vision and Research Agenda for Extending the Reach of Enterprise Modelling, Business and Information Systems Engineering, Volume 60, Issue 1, pp 69-80, Springer.
2 Fill, Hans-Georg (2009): Visualisation for Semantic Information Systems, Gabler.
3 Bork, Domenik, Fill, Hans-Georg (2014): Formal Aspects of Enterprise Modeling Methods: A Comparison Framework, Proceedings of the 2014 47th International Conference on System Sciences, IEEE.
4 Fill, Hans-Georg, Höfferer, Peter (2006): Visual Enhancements of Enterprise Models, in: Lehner, F., Nösekabel, H., Kleinschmidt, P. (2006): Multikonferenz Wirtschaftsinformatik 2006, GITO Verlag, 541-550.

## 3.10 Supporting and Assisting the Execution of LooseLy Framed and Knowledge-intensive Processes

*Frederik Gailly (Ghent University, BE)*



Modeling loosely framed and knowledge-intensive business processes with the currently available process modeling languages is very challenging. Some lack the flexibility to model this type of processes, while others are missing one or more perspectives needed to add the necessary level of detail to the models. In this project we have composed a list of requirements that a modeling language should fulfil in order to adequately support the modeling of this type of processes. Based on these requirements, a metamodel for a new modeling language was developed that satisfies them all. The new language, called DeciClare, incorporates parts of several existing modeling languages, integrating them with new solutions to requirements that had not yet been met. Deciclare is a declarative modeling language at its core, and therefore, can inherently deal with the flexibility required to model loosely framed processes. The complementary resource and data perspectives add the capability to reason about, respectively, resources and data values. The latter makes it possible to encapsulate the knowledge that governs the process flow by offering support for decision modeling. The abstract syntax of DeciClare has been implemented in the form of an Ecore model. In order to also make it possible to automatically discover a DeciClare model we also developed DeciClare Miner. Currently both the language and the miner are evaluated in an Emergency Department of a Belgian Hospital.

## 3.11 Achieving resilience and robustness in strategic models

*Aditya K. Ghose (University of Wollongong, AU)*



The need to future-proof businesses is widely acknowledged as one of the hardest challenges facing business decision makers. Businesses need to anticipate market movements, price movements, regulatory/legislative changes and the likely behaviour of competitors. Much of what happens in the business environment (the effects of moves by these actors) is *adversarial* in nature. *Strategic resilience* requires that businesses make decisions that are most resilient

to adversarial moves by players in the business environment. We cast the strategic resilience problem in the context of organizational goal models. Specifically, we address the problem of selecting the most resilient alternative means of realizing a goal/strategy. We offer a novel means of supporting this decision by using game tree search. We offer a novel data structure that leverages the notion of state update drawn from the literature on reasoning about action, over which we apply game tree search. We show that MINIMAX search and Monte Carlo Tree Search can both underpin a machinery that scales and that makes solving problems of sizes commonly encountered in real-life decision-making feasible.

In this talk, I will also argue that very similar intuitions can underpin flexible business process execution.

I will also provide a brief preview of my Friday talk.

## 3.12 Liberating Modelers from the Tyranny of a Strict Modeling Language

*Martin Glinz (Universität Zürich, CH)*


**Joint work of** Dustin Wüest, Norbert Seyff, Martin Glinz
**Main reference** Dustin Wüest, Norbert Seyff, Martin Glinz: "FlexiSketch: a lightweight sketching and metamodeling approach for end-users", in Software and Systems Modeling, pp. 1–29, Springer, 2017.
**URL** https://doi.org/10.1007/s10270-017-0623-8

Classic modeling tools do not work well in situations where modelers want to sketch ideas, without being constrained by the syntax of a modeling language. This is, for example, the case in creative requirements elicitation and design sessions. On the other hand, whiteboards or paper provide the required flexility of using any kind of notations, but the resulting diagram sketches are just uninterpreted drawings which do not have any syntactic or semantic information associated with the drawn elements. So there is nothing that could be exploited for interpreting the sketches or evolving them into models that could be further processed in a classic modeling tool.

What would be needed is a lightweight modeling approach that supports freeform sketching, but also lets the modeler assign meanings to the drawn elements, thus enabling (a) stepwise transformation from sketches into semi-formal models, and (b) the co-evolution of models and their metamodels.

FlexiSketch, which has been developed in the Requirements Engineering Research Group at the Department of Informatics of the University of Zurich, is a tool providing exactly these capabilities. It is a mobile tool for model-based sketching of free-form diagrams that allows the definition and re-use of diagramming notations on the fly. FlexiSketch lets users draw any node-and-edge diagram they want and recognizes the drawn elements as individual entities and relationships between them. When users assign types and further meta-information to the drawn elements, FlexiSketch generates a lightweight metamodel in the background. FlexiSketch thus supports the co-evolution of models and metamodels. Both models and metamodels can be exported as XML files and then be used for further processing in other tools.

The latest version of the tool, called FlexiSketch TEAM, also supports collaboration with multiple tablets and an electronic whiteboard, such that several users can work simultaneously on the same model sketch.

More information about FlexiSketch is available at http://www.flexisketch.org.

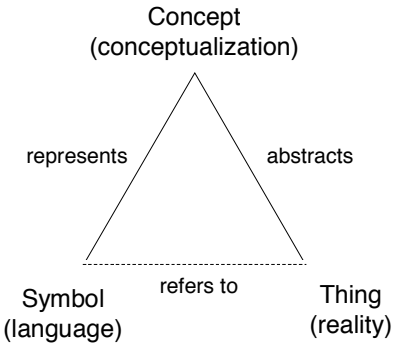



**Figure 1** The Semiotic Triangle.

## 3.13 Yet the Same Look at Models

*Giancarlo Guizzardi (Free University of Bozen-Bolzano, IT)*



In his 1967 classic paper "Another Look at Data" [1], Mealy brought, perhaps for the first time to the attention of this community, the Semiotic Triangle connecting Reality, Conceptualizations and Symbolic Representations (see Figure 1). As he reminds us there, the latter are representations of Conceptualizations. In other words, the relation between Representations and Reality is always mediated by a Conceptualization. Moreover, Mealy reminds us that "data are fragments of a theory of the real-world" and that this is an issue of "Ontology, or the question of what exists". In line with this view, the first point I defend in this talk is that concepts are a prerequisite for the existence of facts, i.e., facts are not in reality but are carved out of reality according to a Reference Conceptualization. In other words, without fixing an a priori conceptualization, there are no determinate Facts! (and, hence, also no Counterfacts).

This view can be depicted in Figure 2, which can be seen as an extension of (one of the sides of) the Semiotic Triangle. As the figure shows, models are representations of abstractions that are carved out of reality according to a certain conceptualization. Moreover, models are grammatically valid constructions built in a modeling language. A language delimits the set of grammatically valid expressions that can be built in that language, in a manner that is analogous to how a conceptualization delimits a set of abstractions (of reality) that it deems acceptable. So, a second point I illustrate in this talk is that the quality of a language to model a set of phenomena in reality can be evaluated and (re)designed by systematically comparing language and conceptualization in these two levels of the Figure 2. To put it simply, a language should contain exactly those modeling primitives that represent the conceptual distinctions put forth by a conceptualization, and it should contain (semantically motivated) syntactical constraints that delimit its set of grammatical models to exactly those that are deemed acceptable by the conceptualization [2, 3]. Finally, as discussed in depth in [3], I defend that systematically analyzing and engineering conceptualizations as in these figures is indeed an issue of "Ontology, or the question of what exists".

### References

1 Mealy, G. H., Another Look at Data, AFIPS Conference Proceedings, Volume 31, Washington, DC: Thompson Books, London: Academic Press, 525–534, 1967.

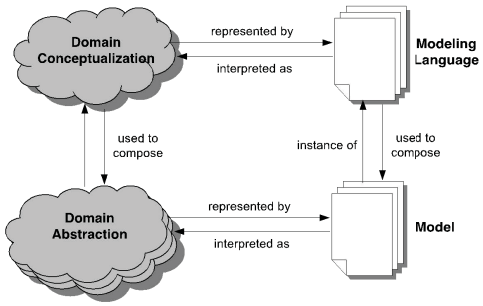



**Figure 2** Conceptualizations and their Abstractions, Modeling Languages and their Models [2, 3].

2 Guizzardi, G., On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models, Frontiers in Artificial Intelligence and Applications, Databases and Information Systems IV, Olegas Vasilecas, Johan Edler, Albertas Caplinskas (Editors), ISBN 978-1-58603-640-8, IOS Press, Amsterdam, 2007.
3 Guizzardi, G.: Ontology-Based Evaluation and Design of Visual Conceptual Modeling Languages, Research Directions in Domain Engineering, Iris Reinhartz-Berger, Arnon Sturm, Tony Clark, Jorn Bettin, Sholom Cohen (Editors), Springer-Verlag, 2013.

## 3.14 Meta Aspects of Operational Conceptual Modeling for Complex Evolving Requirements

*Kamalakar Karlapalem (IIIT – Hyderabad, IN)*



The key idea is to question what does the model do? If it is only an understanding artifact then how to apply the understanding and where. This, we explore the interplay between comprehensible and non-formal expressive conceptual models along with the enactment of the software solution underneath. We present examples from e-contracts and smart solutions to extrapolate on this interplay.

### References

1 Himanshu Jain, P. Radha Krishna, and Kamalakar Karlapalem. e-contract enactment system for effective e-governance. In *7th International Conference on Theory and Practice of Electronic Governance, ICEGOV '13, Seoul, Republic of Korea, October 22-25, 2013*, pages 344–345, 2013. doi:10.1145/2591888.2591951.
2 Himanshu Jain, P. Radha Krishna, and Kamalakar Karlapalem. Context-aware workflow execution engine for e-contract enactment. In *Conceptual Modeling – 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings*, pages 293–301, 2016. doi:10.1007/978-3-319-46397-1_23.
3 Anushree Khandekar, P. Radha Krishna, and Kamalakar Karlapalem. A methodology and toolkit for deploying contract documents as e-contracts. In *Challenges in Conceptual Modelling. Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling – ER 2007. Auckland, New Zealand, November 5-9, 2007. Proceedings*, pages 91–96, 2007.

4 P. Radha Krishna and Kamalakar Karlapalem. Active meta modeling support for evolving e-contracts. In *Challenges in Conceptual Modelling. Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling – ER 2007. Auckland, New Zealand, November 5-9, 2007. Proceedings*, pages 103–108, 2007.

5 P. Radha Krishna and Kamalakar Karlapalem. Electronic contracts. *IEEE Internet Computing*, 12(4):60–68, 2008. doi:10.1109/MIC.2008.77.

6 P. Radha Krishna and Kamalakar Karlapalem. Data, control, and process flow modeling for iot driven smart solutions. In *Conceptual Modeling – 36th International Conference, ER 2017, Valencia, Spain, November 6-9, 2017, Proceedings*, pages 419–433, 2017. doi:10.1007/978-3-319-69904-2_32.

7 P. Radha Krishna and Kamalakar Karlapalem. Event-context-feedback control through bridge workflows for smart solutions. In *Conceptual Modeling – 37th International Conference, ER 2018, Xi'an, China, October 22-25, 2018, Proceedings*, pages 626–634, 2018. doi:10.1007/978-3-030-00847-5_46.

8 P. Radha Krishna, Kamalakar Karlapalem, and Dickson K. W. Chiu. An er$^{ec}$ framework for e-contract modeling, enactment and monitoring. *Data Knowl. Eng.*, 51(1):31–58, 2004. doi:10.1016/j.datak.2004.03.006.

9 P. Radha Krishna, Kamalakar Karlapalem, and Ajay R. Dani. From contracts to e-contracts: Modeling and enactment. *Information Technology and Management*, 6(4):363–387, 2005. doi:10.1007/s10799-005-3901-z.

10 P. Radha Krishna, Anushree Khandekar, and Kamalakar Karlapalem. Modeling dynamic relationship types for subsets of entity type instances and across entity types. *Inf. Syst.*, 60:114–126, 2016. doi:10.1016/j.is.2016.03.010.

11 Nishtha Madaan, P. Radha Krishna, and Kamalakar Karlapalem. Consistency detection in e-contract documents. In *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2014, Guimaraes, Portugal, October 27-30, 2014*, pages 267–274, 2014. doi:10.1145/2691195.2691249.

12 Pabitra Mohapatra, P. Radha Krishna, and Kamalakar Karlapalem. E-contract enactment using meta execution workflow. In *On the Move to Meaningful Internet Systems: OTM 2013 Workshops – Confederated International Workshops: OTM Academy, OTM Industry Case Studies Program, ACM, EI2N, ISDE, META4eS, ORM, SeDeS, SINCOM, SMS, and SOMOCO 2013, Graz, Austria, September 9 – 13, 2013, Proceedings*, pages 714–717, 2013. doi:10.1007/978-3-642-41033-8_91.


## 3.15 Challenges in Improving Collaboration in Conceptual Modeling

*Julio Cesar Leite (PUC – Rio de Janeiro, BR)*



**Main reference** Julio Cesar Sampaio do Prado Leite: "The Prevalence of code Over Models: Turning it Around with Transparency", in Proc. of the 8th IEEE International Model-Driven Requirements Engineering Workshop, MoDRE@RE 2018, Banff, AB, Canada, August 20, 2018, pp. 56–57, IEEE Computer Society, 2018.

**URL** https://doi.org/10.1109/MoDRE.2018.00013

Open source software has been very successful since it relies on massive collaboration.

Massive collaboration involves a very large number of programmers working on the same project in different locations in an asynchronous way. However, to achieve this type of collaboration, programmers use a proper infrastructure to support the basic issues of collaboration, which heavy relies on configuration management.

An example of such infrastructure is GitHub[1], with millions of repositories and millions of users. We [2] and others[3] believe that massive collaboration is possible because GitHub helps transparency, thus allowing for this kind of social interaction among programmers.

As modeling is fundamental in software construction/evolution, we posit[4] that we need to tackle the issue of how to bring massive collaboration towards the process of building/evolving conceptual models. Several obstacles do exist. In particular, we believe that three of them are paramount:

**a)** Reuse,
**b)** Transparency, and
**c)** Collaboration mechanisms.

A possible path to Reuse is by means of domain-oriented models/patterns. As for Transparency, the understanding of GitHub mechanics and their application on modeling infrastructures seems a way to proceed. Collaboration mechanisms for conceptual modeling do exist and are being used/studied on site (same location in a synchronous mode), so we should research the adaptation/extension of these mechanisms to modeling infrastructures.

### References

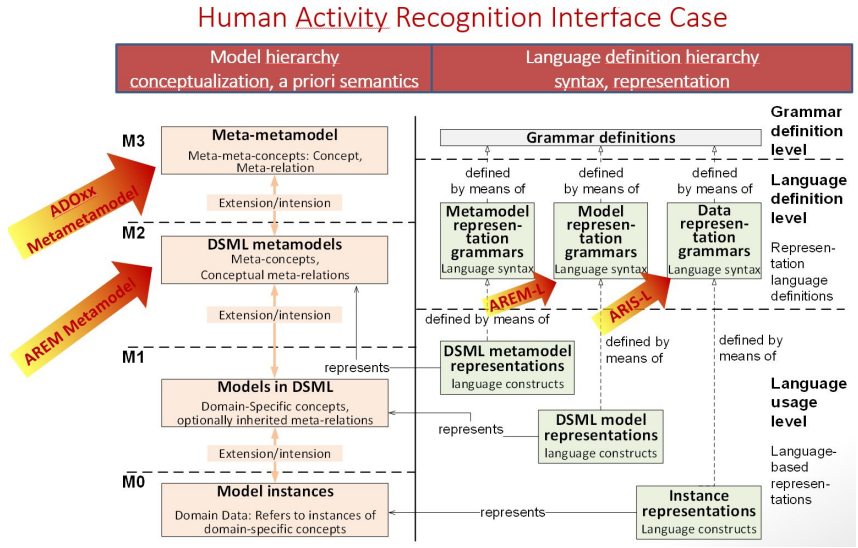
1 https://github.com/
2 J.C.S.P. Leite, C. Cappelli, *Software Transparency*, Bus Inf Syst Eng, vol. 2, pp. 127, 2010.
3 Laura Dabbish, et al. *Social coding in GitHub: transparency and collaboration in an open software repository.* In Proceedings of the CSCW '12. ACM, (2012),1277–1286.
4 Leite, J.C.S.P., *The Prevalence of code Over Models: Turning it Around with Transparency.* MoDRE@RE 2018: 56–57

## 3.16 Extraction and validation of Structural Models by using AI/ML

*Wolfgang Maaß (Universität des Saarlandes, DE)*



Automatic validation of structural models interferes with the deductive research method in information systems research. Nonetheless it is tempting to use a statistical learning method for assessing meaningful relations between structural variables given the underlying measurement model. In this talk, we discuss the epistemological background for this method and describe its general structure. Thereafter this method is applied in a mode of inductive confirmation to an existing data set that has been used for evaluating a deductively derived structural model. In this study, a range of machine learning model classes is used for statistical learning and results are compared with the original model.




## 3.17 The Paradigm of Model Centered Architecture (MCA)

*Heinrich C. Mayr (Alpen-Adria-Universität Klagenfurt, AT)*


**Joint work of** Heinrich C. Mayr, Judith Michael, Suneth Ranasinghe, Vladimir A. Shekhovtsov, Claudia Steinberger
**Main reference** Heinrich C. Mayr, Judith Michael, Suneth Ranasinghe, Vladimir A. Shekhovtsov, Claudia Steinberger: "Model Centered Architecture", in Proc. of the Conceptual Modeling Perspectives., pp. 85–104, Springer, 2017.
**URL** https://doi.org/10.1007/978-3-319-67271-7_7

The MCA paradigm is based on the obvious fact that any type of data managed and/or processed within a digital ecosystem, as well as the processes themselves, are instances of explicitly specified or implicitly underlying models and are thus models again. We therefore see each software and system component as a construct consisting of model handlers (consumers and/or producers). MCA can be seen as a generalization of Model Driven Architecture (MDA), Model Driven Software Development (MDSD) and models@runtime. Like multi-level modeling, MCA advocates the use of (possibly recursive) hierarchies of domain-specific modeling languages (DSML) for any system aspect, each embedded in a suitable methodological framework. Thus, all system interfaces are also defined via models using a corresponding DSML. This means that MCA concentrates on several meta-models and models in each development step up to the running system. The semantic concepts defined by the model hierarchies are to be represented by suitable representation languages, which again form a hierarchy. In the lecture, the MCA paradigm was illustrated with an example from the field of assistive systems: It was shown how arbitrary human activity detection systems can be docked to a support system via a meta-model-based interface specification. In addition, a number of interesting open questions were raised, such as the extension of meta-model frameworks to multi-metamodel environments, the alignment with agile software development process models like SCRUM or the need of mechanisms for meta-model reuse.

### References

1 E. Ströckl and H.C. Mayr. *Multi-Modal Human-System Interaction based on MCA* . Proc. 2nd International Conference on Intelligent Human Systems Integration: Integrating People and Intelligent Systems, Feb. 2019, San Diego, California.

2 V. Shekhovtsov, S. Ranasinghe, H.C. Mayr and J. Michael. *Domain Specific Models as System Links*. Advances in Conceptual Modeling, LNCS 11158, Springer Int., 2018.
3 H.C. Mayr, J. Michael, V. Shekhovtsov, S. Ranasinghe and C. Steinberger. *A Model Centered Perspective on Software-Intensive Systems*. Proc. 9th Int. Workshop on Enterprise Modeling and Information Systems Architectures, Rostock Germany, CEUR-WS.org Vol 2097, pp 58–64.

## 3.18 Conceptual Modeling and MDSE – Two worlds on one planet

*Judith Michael (RWTH Aachen, DE)*


**URL** https://materials.dagstuhl.de/files/18/18471/18471.JudithMichael.Slides.pdf

Every researcher has a certain definition of the main scientific terms in mind. These definitions are based on our socialization, e.g., the research group we made our PhD in, the scientific community where we attend most conferences, the researchers we work and communicate with, the application domains we work on. In communication, we use these terms with respect to our own definitions in mind and our counterpart with his own definitions. Since these backgrounds can vary greatly, discussions on the same topic often run in different directions. For the conceptual modeling community this aspect is even more important as there is not only one main community all researchers belong to: they are e.g., related to databases, software engineering, ontologies, formal methods, petri nets, business processes. These communities even have a variety of main conferences. The exchange between these conferences is low and researchers stay in their filter bubbles. Because of these different backgrounds the understanding of e.g. a project on the semantic difference of models, can be either understood as a project using computational linguistics and ontologies to solve this challenge or on using denotational semantics and mathematical calculations. Thus, terms such as semantics, models, conceptual models, domain or domain specific (modeling) language are interpreted in different ways. To make an improvement of the current situation, we plan to publish our own work on User-Centered and Privacy-Driven System Design, as well as the research in the MaCoCo (Management Cockpit for Chair Controlling and Science Management) and SemanticDiff Projects not only in one community but to show it to different ones. In order to bring the conceptual modeling community closer together it is important to (1) make the different understandings explicit to be able to talk more conscious, (2) establish a platform or institution for regular exchange and (3) very concretely: to know the literature of the different communities and conferences and to publish the own research more widely. These aspects will be discussed even more intensively during this Dagstuhl Seminar.

## 3.19 Using High-level Petri Nets in Domain Specific Language Design

*Daniel Moldt (Universität Hamburg, DE)*



Creating DSLs for a domain or project is complex if you need a tool to draw models in addition to language. If models are run with the tool for simulation or animation purposes, the challenges become even greater. General solutions are currently not available.

Equipped with our Renew tool (http://www.renew.de) and our RMT framework (Renew Meta Modeling and Transformation), we address all kinds of dynamic models. States, state changes / transitions, events, processes and related terms (firing, activation, conflict resolution, synchronization etc.) can be covered with RMT in the development of DSLs. Thus, we offer transformative semantics to provide an underlying Petri net design for such DSLs.

This contribution focuses on the provision of simulation and animation feedback for models of such DSLs. For such DSLs, all common concepts of Petri nets are provided. In this way, we can highlight the desired properties for graphical elements in a model as desired by DSL model users. Activation, triggering, synchronization, conflicts, concurrency etc. can be covered.

Based on an abstract syntax, a concrete syntax and a tool configuration, meta-modeling offers the possibility to feed the RMT framework in such a way that a tool is generated. As stated already, the DSL tool supports the modeling of DSL models and their simulation and visualization.

The entire approach is illustrated by some examples from the area of Business Process Model and Notation (BPMN).

## 3.20 Conceptual modeling for Social networks and Crowdsourcing to support emergency management

*Barbara Pernici (Polytechnic University of Milan, IT)*



Developing complex information systems for specific domains is becoming more and more frequent and the stakeholders require a good understanding of the domain. For conceptual modelers it is difficult to move from one domain to another, so a possible way is to become a conceptual modeler expert in a given domain. Even in this case, it will always be needed to evolve models as new requirements arise. In this paper we present the approach developed in the European H2020 E2mC project on Evolution of Emergency Management Services in Copernicus and we discuss how data and services have been modeled. Some emerging concepts for new conceptual modeling methods are discussed in the presentation, in the direction of making the models closer to the users' and their data. An initial proposal in the direction of providing support to domain experts for designing and revising their own models in their own terms is discussed.

## 3.21 Contextual Aspects in Situational Method Engineering

*Jolita Ralyté (University of Geneva, CH)*


**Joint work of** Jolita Ralyté, Xavier Franch
**Main reference** Jolita Ralyté, Xavier Franch: "Using Contextual Goal Models for Constructing Situational Methods", in Proc. of the Conceptual Modeling – 37th International Conference, ER 2018, Xi'an, China, October 22-25, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 11157, pp. 440–448, Springer, 2018.
**URL** https://doi.org/10.1007/978-3-030-00847-5_31

Situational method engineering (SME) has emerged as a result of the common recognition that one-size-fits-all methods can never be totally successful in a constantly changing information systems development (ISD) environment. The mission of SME consists in providing concepts and guidance for situation-specific (i.e., situational) method construction and adaptation by reusing various types of method chunks. Many approaches have been proposed by now, but the contextual aspects of SME are still a subject for investigation and formalization. Indeed, situation and intention are two fundamental notions in SME. They are used to assess the situation of an ISD project and to specify method requirements in this situation. They also allow defining the goals of the method chunks and the conditions under which they can be applied. In this way, the selection and assembly of method chunks for a particular ISD project is driven by matching situational method requirements to method chunks' goals and context descriptions. In our current work we propose to use contextual goal models for dealing with intentional and contextual aspects in SME, and even supporting all SME steps. Our approach is based on iStar2.0 modeling language that we extend with contextual annotations.

## 3.22 The quest for a general framework for composition and compositionality of conceptual models

*Wolfgang Reisig (HU Berlin, DE)*


**Main reference** Wolfgang Reisig: "Associative composition of components with double-sided interfaces", Acta Inf., Vol. 56(3), pp. 229–253, 2019.
**URL** https://doi.org/10.1007/s00236-018-0328-7

Informatics is nowadays about large systems that consist of heterogeneous components, including not only software packages, but also people, things, services, etc. Composition of such components is an essential issue. A general, fundamental, theory driven, formal basis to systematically compose components, would decisively improve systematic design of nowadays large informatics systems. This talk presents a number of requirements at such a formal basis, and suggests a formal framework that fits these requirements. The essential idea is to be liberal on the description of the inner behavior of components, but to be strict on the technicalities of composition. In particular, associativity of composition must be guaranteed. It is shown that the suggested framework adequately covers various different examples and case studies.

## 3.23 Traceability Engineering: A research agenda

*Marcela Ruiz Carmona (Utrecht University, NL)*



The larger the software or systems development project, the more engineers and engineering artefacts – and thereby traceability links – are involved. Large-scale software or systems development can involve many thousands artefacts from heterogeneous systems like source code, test cases, requirements, crowd-sourced feedback, and elicited data. In the context of governmental, healthcare, and financial institutions, the role of traceability is highly appreciated but not exploited.

In the context of this Dagstuhl seminar, I share my research agenda for Traceability Engineering. My six-year research goal is to establish the Traceability Engineering Lab. I see great potential on combining current Big Data and IoT technologies for the management and exploitation of traceability in large-scale software and information systems development projects. The three pillars of the lab are presented, which motivated the attendees to provide valuable feedback in terms of quantitative metrics, visualisation challenges, the opportunities to bring conceptual modelling for agile software development, and the need for formal definitions.

## 3.24 Engineering Software Languages

*Bernhard Rumpe (RWTH Aachen, DE)*



We discuss about the "Engineering" aspect in the engineeering of software languages. In particular, composition of models, refinement and a good notion of modularity are important.

Heterogeneous languages, such as SysML, however, lift the notion of model composition also to the notion of language composition, which needs to be understood in detail to be able to use models for describing different abstractions and aspects of a system (or a "world").

Many more aspects are related to SLE:
http://www.se-rwth.de/topics/Language-Engineering.php

## 3.25 Conceptual modelling of real-time and real-space aspects for cyber-physical systems and processes

*Heinz W. Schmidt (RMIT University – Melbourne, AU)*



My talk summarises some lessons and challenges from a few past research projects in conceptual and architectural modelling of cyberphysical systems (CPS) and software. These projects were focused on extra-functional properties, model-based verification and testing, but also on collaboration platform architecture for widely distributed multi-disciplinary design and development teams.

The term internet of people, things and services (IoPTS) was coined to stress the ultra-large scale character of such systems including (for some vendors) extremely large numbers of human actors, devices and services. Modeling, analysis and architectural design are not limited to software and data in these systems of systems, but include organisational, human and artificially intelligent actors, as well as very large and very small physical systems and processes.

Fur our industry collaborators, the physical aspects related to robotics automation, remote management of plants or computational and physical science experiments. Many challenges remain. Not the least of these is that CPS cross boundaries of professions and expertise, for example mechatronics engineers, software engineers, business analysts, computer scientists, industrial designers and others. Each have significantly different foundations, standards and practices in modelling, analysis and design.

Consequently architectural notations (typically based on formalising annotations on top of 'boxes and lines' drawings) remain of great interest to us. For, architectural language is the lingua franca connecting the different disciplines. While much progress has been made in modeling architecture over the past couple of decades, the composition of different formal and informal domain-specific models associated with elements in the shared architecture however remains elusive.

## 3.26 Analytical Patterns: Domain-independent and domain-specific cores of analytical queries

*Michael Schrefl (Johannes Kepler Universität Linz, AT)*



Analytical query patterns capture the reusable core of analytical queries. They are in business analytics the counterpart to design patterns in software engineering.

An analytical pattern is defined by (1) a set of pattern elements, (2) a set of constraints over pattern elements, (3) an a pattern expression with pattern elements embedded in some analytical query language such as SQL. Each pattern element is a named placeholder of one or a list of element(s) of the dimensional-fact-model (DFM) of a data warehouse, such as dimension, level, fact, or measure. Pattern elements may be input-parameters, result-parameters, or local pattern elements. Analytical patterns are best exploited by using an enriched the DFM-model that comes with ontologies of predicates (over facts or dimensions) and of calculated measures that can also constitute elements of an analytical pattern.

An analytical pattern is practically/fully instantiated by binding some/each formal input parameter element either to the name for a DFM-element (to be bound later during application) or to the identity, e.g., URI, of a DFM-element (static binding). A fully instantiated analytical pattern is applied to a specific data warehouse (DWH) by identifying the data warehouse context and by dynamically binding the name of actual pattern elements to the identity of a DFM-element in the indicated DWH-context. A pattern application is valid if the bindings of the pattern elements satisfy the pattern constraints.

We present a set of domain-independent analytical patterns identified by generalizing similar analytical queries frequently accounted in various domains (such as medicine, farming, and production). We present selected domain-independent analytical patterns that are defined by partially instantiating domain-independent patterns with facts, dimensions, levels, predicates, and calculated measure of the domain ontology, whereby bound input elements become local elements.

## 3.27 Recitals by computer scientists

*Sibylle Schupp (TU Hamburg, DE)*



The General Data Protection Regulation (GDPR) defines "Privacy by Design" as "data protection through technology design." Other privacy laws or regulations have a similar view so that it is common for legal texts in this domain to contain several references to "technology" or "the state of the (technological) art." Formal methods are suited to transform legal wordings into "echnology." Unless one is an expert in formal methods, however, it is not always obvious which legal terms are subject to a formal specification, and in what way, nor, conversely, which legal terms are underspecified and cannot be formalized without additional assumptions. For questions on legal interpretations, one can resort to recitals, which are associated with particular articles and provide additional explanations. What could be the counterpart for open issues concerning the formalization of a privacy regulation?

## 3.28 Realizing Digital Ecosystems in MCA

*Vladimir Shekhovtsov (CICERO Consulting GmbH – Klagenfurt, AT)*


**Joint work of** Heinrich C. Mayr, Judith Michael, Suneth Ranasinghe, Vladimir A. Shekhovtsov, Claudia Steinberger

I presented the realization of digital ecosystems in Model Centered Architecture (MCA) on an example of Ambient Assistance system. This example is based on the Human Behavior Monitoring System (HBMS) project which goal is to preserve the individual episodic memory of a person by building cognitive model of his/her behavior, and to exploit this model for support (ambient assistance) in case of cognitive impairments. The language architecture of the HBMS system features four different modeling languages on M1 level, and some more M0 representation languages, this architecture can be exemplified by the metamodel-based Human Cognition Modeling Language (HCM-L) which is used to describe the models of human behavior and the necessary contexts, to be stored and exploited by the system according to the MCA paradigm. The HBMS system architecture is built as an instance of MCA; it includes the following components: (1) the modeling tool implemented by means of ADOxx metamodeling framework, (2) the system kernel (3) the model transfer interface connecting (1) and (2), (4) the model storage, (5) the interface to external Human Activity Recognition (HAR) systems implemented as a set of MCA links. The user's behavior is captured by HAR systems by means of sensors, then the recognized structures are transferred to the kernel, where they are matched against the structure of stored behavioral models of the supported person to find possible inconsistencies, made predictions, and provide support. The support is provided by means of multimodal user support interface featuring audio and visual output. The system kernel also provides the monitoring interface which uses the graphical representation of the models to provide information about current position of the user within the behavioral scenario, and the simulator for sensor output. The HBMS system was validated by implementing it as a part of the sensor lab, which was then used to test its functionality to support human participants.

## 3.29 Model-based analysis of runtime business process behavior

*Pnina Soffer (Haifa University, IL)*


**Joint work of** Pnina Soffer, Yotam Evron, Arava Tsoury, Anna Zamansky, Iris Reinhartz-Berger

Conceptual process models are typically considered as design-time artifacts, aimed at establishing an understanding of a business process, redesigning it, or communicating about it. The talk presented three approaches which use process models as a basis for analyzing aspects of process runtime behavior. First, an analysis approach of potential data quality problems that may occur at runtime. The approach is based on an ontology-based view of a process, and a formal notion of Data-Inaccuracy-Awareness (DIA), which indicates that at a given state in a process, it is known whether data values accurately reflect real world values. An algorithm was developed for automatically analyzing a process model and identifying where data is used by the process at non-DIA states. With this analysis it is possible to add controls to the process (at design time) and avoid data inaccuracy consequences during execution. Second, a notion of cross-instance data impacts, which relates to variables / data items that are shared by different process instances (e.g., resource capacity). Events and changes in such variables, that take place in one process instance, might affect the state of other process instances. However, such effects are apparent in process models, which typically depict a single process instance. The proposed approach analyzes the cross-instance impacts of unexpected changes in values of variables and identifies where responses are needed. Last, we propose a notion of conditional conformance checking, which extends existing conformance checking techniques between an actual (executed) business process and the (prescribed) process model. Once an unexpected deviation from the model takes place, it is expected that some responses and compensations will take place rather than that the process would continue as if nothing happened. Our conditional conformance takes this into account. Upon an unexpected deviation, the expected response is automatically calculated (based on the model). The conformance measurement relates to the "normative" process model as well as to the expected compensating actions that should follow a deviation.

## 3.30 Lessons learnt from the design and development of a method and domain-specific language for security-risk assessment – The CORAS experience

*Ketil Stølen (SINTEF – Oslo, NO)*



This talk presents lessons learnt from the design and development of the CORAS approach for security risk assessment. The work on CORAS was initiated in 2001 and reached a major milestone in 2015 with the publication of the CORAS book. The talk starts by giving a brief overview of CORAS with particular focus on threat modelling. We then go on to present our experiences and what we learnt. Finally, we try to align our work with the overall Dagstuhl-theme of languages for conceptual modeling.

## 3.31 Deep and Normal Models

*Bernhard Thalheim (Universität Kiel, DE)*


**Joint work of** the Kiel model-to_model-modelling MMM team
**Main reference** Bernhard Thalheim: "Normal Models and Their Modelling Matrix", in Proc. of the Models: Concepts, Theory, Logic, Reasoning and Semantics – Essays Dedicated to Klaus-Dieter Schewe on the Occasion of his 60th Birthday, pp. 44–74, College Publications, 2018.

There are many notions of the (conceptual) model. One of them is the following general one. The problem is, however, whether all facets of this notion are essential within the given application and utilisation scenario, for the given community of practice, within the given context, for the current scope and focus of interest, and the profile of the model. In most case, we may restrict ourselves to some of them an thus develop "normal" models. The rest of the model is inherited from the "deep" model that is shuffled into the normal model and thus form the foundation of the normal model.

A model is a well-formed, adequate, and dependable instrument that represents origins and functions in some utilisation scenario. A model is a representation of some origins and may consist of many expressions such as sentences. Adequacy is based on satisfaction of the purpose or function or goal, analogy to the origins it represents and the focus under which the model is used. Dependability is based on a justification for its usage as a model and on a quality certificate. Models can be evaluated by one of the evaluation frameworks.

A model is functional if methods for its development and for its deployment are given. A model is effective if it can be deployed according to its portfolio, i.e. according to the tasks assigned to the model. Deployment is often using some deployment macro-model, e.g. for explanation, exploration, construction, documentation, description and prescription.

Models function as instruments or tools. Typically, instruments come in a variety of forms and fulfill many different functions. Instruments are partially independent or autonomous of the thing they operate on.

## 3.32 Automatic Experiment Generation for Supporting the Analysis of Domain Specific Simulation Models

*Pia Wilsdorf (Universität Rostock, DE)*


**Joint work of** Pia Wilsdorf, Andreas Ruscheinski, Kai Budde, Tom Warnke, Bjarne Christian Hiller, Marcus Dombrowsky, Adelinde M. Uhrmacher

Domain-specific modeling approaches, such as ML-Rules [1], play an important role in modeling biological systems since they are able to capture the complex dynamics between multiple levels of organization. The development and analysis of such models involves a wide variety of simulation experiments. In recent years, domain-specific languages (e.g., SESSL [2]) have also been applied for expressing experiment specifications, thereby making this part of a simulation study explicit and easier to replicate. However, writing such specifications can be challenging even if a specification language exists, as questions about models and the experiments necessary to answer them are becoming more complex and diverse. Therefore, to facilitate the specification of simulation experiments we create templates for certain experiment types, such as sensitivity analysis, or statistical model checking, which can then

be adapted to the concrete simulation model and study based on information provided in the model's documentation [3]. We propose an automatic extraction procedure, however, the lack of (semi-) formal documentations hinders a fully-automatic extraction. Furthermore, the lack of an explicit conceptualization for simulation experiments and their constituents makes a context-dependent inference of information difficult.

### References

1 Maus, Carsten and Rybacki, Stefan and Uhrmacher, Adelinde M. (2011) Rule-based multi-level modeling of cell biological systems. BMC Systems Biology, 5(1), p.166.
2 Warnke, Tom and Helms, Tobias and Uhrmacher, Adelinde M. (2018) Reproducible and flexible simulation experiments with ML-Rules and SESSL. Bioinformatics, 34 (8), pp. 1424–1427
3 Ruscheinski, Andreas and Budde, Kai and Warnke, Tom and Wilsdorf, Pia and Hiller, Bjarne Christian and Dombrowsky, Marcus and Uhrmacher, Adelinde M. (2018) Generating Simulation Experiments Based on Model Documentations and Templates. In: Winter Simulation Conference (WSC 2018), 09-12 Dec 2018, Gothenburg, Sweden.

## 3.33 Modeling for Industry 4.0

*Manuel Wimmer (TU Wien, AT)*


**Joint work of** Luca Berardinelli, Stefan Biffl, Emanuel Mätzler, Tanja Mayerhofer
**Main reference** Luca Berardinelli, Stefan Biffl, Emanuel Mätzler, Tanja Mayerhofer, Manuel Wimmer: "Model-based co-evolution of production systems and their libraries with AutomationML", in Proc. of the 20th IEEE Conference on Emerging Technologies & Factory Automation, ETFA 2015, Luxembourg, September 8-11, 2015, pp. 1–8, IEEE, 2015.
**URL** https://doi.org/10.1109/ETFA.2015.7301483

Production systems are becoming more and more software-intensive, thus turning into cyber-physical production systems (CPPS). This is also highlighted and reflected by Industry 4.0, which is seen as the next industrial revolution. As with the previous industrial revolutions, new demands have to be satisfied, e.g., virtually exploring variants, finding optimal solutions, and making dynamic runtime decisions, to allow companies to be more competitive. As a consequence, however, the complexity of CPPS is increasing. To deal with this increased complexity, modeling is considered as a promising approach which is explored in several academic and industrial efforts.

In my talk, I will introduce one of the most prominent family of modelling languages in the context of Industry 4.0, namely AutomationML (www.automationml.org). In particular, I will present some lessons learned from past and ongoing projects dealing with AutomationML and outline challenges and opportunities in this realm for the conceptual modelling research community. To sum up, the main open research question is how to best reach engineers in different engineering disciplines?

# 4 Working groups

## 4.1 Working group on Grand Challenges in Conceptual Modeling

*João Paulo Almeida (Federal University of Espírito Santo – Vitória, BR),João Araújo (New University of Lisbon, PT), Fernanda Baião (PUC – Rio de Janeiro, BR), Giancarlo Guizzardi (Free University of Bozen-Bolzano, IT), and Pnina Soffer (Haifa University, IL)*



Here we report on working group discussions that were driven by five questions posed by the seminar organizers. The members of this break-out group were João Paulo A. Almeida (responsible for these notes), João Araújo, Fernanda Baião, Giancarlo Guizzardi and Pnina Soffer.

1. *Your university plans to establish an Institute of Modeling; one research group including a full professorship of that institute is planned for conceptual modeling research and teaching. What should the job posting look like?*
   Capable of leading high-quality research on theories, methods and tools for producing cognitively-effective conceptual models that convey real-world semantics. Experience with specific high-impact application domains a plus.
2. *Suppose a donator gives you 5 million dollars for research in conceptual modeling: what would you be researching?*
   We have identified two options:
   a. Establish a Network-of-Excellence on Conceptual Modeling, or;
   b. Run a Research Project on Foundations and Applications, which would emphasize interdisciplinarity and would have an exploratory character. It would entail:
      i. Investigating foundations;
      ii. Identifying high priority domains, and;
      iii. Selecting complex (high-impact) problems for conceptual modeling for experimentation.
3. *How can we make 1 and 2 happen?*
   Frame (or disguise `:-)`) the work as a machine learning project. Joke aside, we have identified that there is wide potential for research into the synergy between conceptual modeling and machine learning. Given the momentum for the latter, there could be funding opportunities in this unexplored intersection.
4. *Which are the most important findings in the CM discipline within the last 10 years?*
   We discussed a number of high-impact results in the last 10 years, but could not identify "major breakthroughs". We concluded we should look back further, and identified the work of Nicola Guarino in the 90s, Ron Weber and Yair Wand in the late 80s, and also Bill Kent in the late 70s.
5. *What do you expect from a conference on Conceptual Modeling in General?*
   Address grand challenges. Accept vision-oriented papers.

## Participants

- João Paulo Almeida
Federal University of Espírito Santo – Vitória, BR
- João Araújo
New University of Lisbon, PT
- Fernanda Baião
PUC – Rio de Janeiro, BR
- Ladjel Bellatreche
ENSMA – Chasseneuil, FR
- Dominik Bork
Universität Wien, AT
- Isabelle Comyn-Wattiau
ESSEC Business School – Cergy Pontoise, FR
- Rocco De Nicola
IMT – Lucca, IT
- Felicita Di Giandomenico
CNR – Pisa, IT
- Hans-Georg Fill
University of Fribourg, CH
- Frederik Gailly
Ghent University, BE
- Aditya K. Ghose
University of Wollongong, AU
- Martin Glinz
Universität Zürich, CH
- Giancarlo Guizzardi
Free University of Bozen-Bolzano, IT
- Kamalakar Karlapalem
IIIT – Hyderabad, IN
- Julio Cesar Leite
PUC – Rio de Janeiro, BR
- Stephen W. Liddle
Brigham Young University, US
- Wolfgang Maaß
Universität des Saarlandes, DE
- Heinrich C. Mayr
Alpen-Adria-Universität Klagenfurt, AT
- Judith Michael
RWTH Aachen, DE
- Daniel Moldt
Universität Hamburg, DE
- Oscar Pastor Lopez
Technical University of Valencia, ES
- Barbara Pernici
Polytechnic University of Milan, IT
- Jolita Ralyte
University of Geneva, CH
- Wolfgang Reisig
HU Berlin, DE
- Marcela Ruiz Carmona
Utrecht University, NL
- Bernhard Rumpe
RWTH Aachen, DE
- Heinz W. Schmidt
RMIT University – Melbourne, AU
- Michael Schrefl
Johannes Kepler Universität Linz, AT
- Sibylle Schupp
TU Hamburg, DE
- Vladimir Shekhovtsov
CICERO Consulting GmbH – Klagenfurt, AT
- Pnina Soffer
Haifa University, IL
- Markus Stumptner
University of South Australia – Adelaide, AU
- Ketil Stølen
SINTEF – Oslo, NO
- Bernhard Thalheim
Universität Kiel, DE
- Pia Wilsdorf
Universität Rostock, DE
- Manuel Wimmer
TU Wien, AT

Report from Dagstuhl Perspectives Workshop 18472

# Implementing FAIR Data Infrastructures

**Edited by**

**Natalia Manola[1], Peter Mutschke[2], Guido Scherp[3], Klaus Tochtermann[4], and Peter Wittenburg[5]**

1 **University of Athens, GR,** natalia@di.uoa.gr
2 **GESIS – Leibniz Institute for the Social Sciences – Cologne, DE,** peter.mutschke@gesis.org
3 **ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE,** g.scherp@zbw.eu
4 **ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE,** k.tochtermann@zbw.eu
5 **Max Planck Computing and Data Facility – Garching, DE,** peter.wittenburg@mpi.nl

## Abstract

This report documents the programme and the outcomes of Dagstuhl Perspectives Workshop 18472 "Implementing FAIR Data Infrastructures". The workshop aimed at bringing together computer scientists with digital infrastructure experts from different domains to discuss open issues implementing and adopting the FAIR principles in research data infrastructures and to shape the role that the field of computer science has to play.

**Seminar** November 18–21, 2018 – http://www.dagstuhl.de/18472
**2012 ACM Subject Classification** Information systems
**Keywords and phrases** fair principles, open data, open science, research data infrastructures
**Digital Object Identifier** 10.4230/DagRep.8.11.91

## 1 Executive Summary

*Natalia Manola (University of Athens, GR)*
*Peter Mutschke (GESIS – Leibniz Institute for the Social Sciences – Cologne, DE)*
*Guido Scherp (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE)*
*Klaus Tochtermann (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE)*
*Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE)*



The Dagstuhl Perspectives Workshop on "Implementing FAIR Data Infrastructure" aimed at bringing together computer scientists and digital infrastructure experts from different domains to discuss challenges, open issues, and technical approaches for implementing the so-called FAIR Data Principles in research data infrastructures. Moreover, the workshop aimed to shape the role of and to develop a vision for computer science for the next years in this field, and to work out the potentials of computer science in advancing Open Science practices.

In the context of Open Science, and the European Open Science Cloud (EOSC) in particular, the FAIR principles seem to become a common and widely accepted conceptual basis for future research data infrastructures. The principles consist of the four core facets that data must be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable in order to advance the


Implementing FAIR Data Infrastructures, *Dagstuhl Reports*, Vol. 8, Issue 11, pp. 91–111
Editors: Natalia Manola, Peter Mutschke, Guido Scherp, Klaus Tochtermann, and Peter Wittenburg
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

discoverability, reuse and reproducibility of research results. However, the FAIR principles are neither a specific standard nor do they suggest specific technologies or implementations. They describe the core characteristics of data use. Thus, the FAIR principles cover a broad range of implementation solutions. This certainly incorporates the risk of having a highly fragmented set of solutions at the end of the day.

Given this, and in view of the "need for a fast track implementation initiative [of the EOSC]"[1], it is strongly needed to turn the principles into practice. Therefore, the workshop took the recommendations of the European Commission Expert Group on FAIR Data "Turning FAIR into reality" as a starting point and discussed what can be done next from the perspective of computer science to enable data providers to make their data FAIR.

The workshop started with three ignition talks on the wider background and context of the FAIR principles (given by Peter Wittenburg), the relationship of FAIR to Open Data (given by Natalia Manola) and the role of the principles within the EOSC (given by Klaus Tochtermann). Based on these talks as well as inputs from all participants in the forefront of the workshop, we have split the discussion into three working groups addressing, for each of the four principles, the main key challenges for implementing FAIR and the question what and how computer science can contribute to these key challenges. Based on the results of these three initial working groups we furthermore split into more focused groups addressing the problem of licenses w.r.t. data use, (self)improvement of FAIRification, and the relation of FAIR and data intensive science.

Finally, we identified three major areas to be addressed in the manifesto which we discussed in three further working groups:

1. Infrastructures & Services Aspects: This group focused on the question by which technical means research data infrastructures and data services can be advanced to better address and fulfil the FAIR principles.
2. Computer Science Research Topics: The working group discussed the relationship of research areas in computer science and topics relevant to implement FAIR data infrastructures.
3. FAIR Computer Science Research: While the other two groups mainly focused on the contribution of computer science to implement FAIR, this working group addressed the question how the FAIR principles are currently adopted by computer science research itself and what should be improved.

The participants will continue their work in the aforementioned issues, and a manifesto is foreseen to be ready by mid May 2019.

[1] https://www.dtls.nl/wp-content/uploads/2017/05/DE-NL-Joint-Paper-FINAL.pdf

## 2 Table of Contents

# 3 Ignition Talks

## 3.1 Why FAIR and what is the context of it?

*Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE)*



Since about 2005, the special nature and value of data was commonly recognised. Since then intensive discussions took place in data science. OECD, a high level group of the EC, a workshop at ICRI 2012 that led to the establishment of the Research Data Alliance (RDA), the group of G8 Science Ministers, various RDA working groups, and many other places with the goal to identify ways to improve data sharing and reuse. These are currently hampered by a huge fragmentation and resulting in inefficiencies and costs. The FAIR principles finally formulated now widely accepted agreements on a minimal set of behaviours about the creation and management of digital objects in a convincing way. They formulate clear messages to change our current data practices which are characterised by 80% of waste of time of data professionals in data projects leading to consequences such as failed projects and exclusion of many experts and SMEs, etc. However, the FAIR principles are not blueprints to build the urgently needed infrastructures that will help changing practices. Initiatives such as RDA[2], GEDE[3], DONA[4] and GO FAIR[5] are now in agreement that the concept of "Digital Objects" (DO), which have a bit sequence encoding some content and are associated with a persistent and unique identifier and different kinds of metadata, is a way to implement interoperability at data organisation and modelling layer. The recently developed DO Interface Protocol (DOIP) therefore is a uniform interface to access all DOs in repositories independent of how these are set-up and the content the DOs are encoding. The FAIR principles and the DO implementation concept are therefore complementary and thus can be pillars of convergence towards improved efficiency in data management and reuse as requested by Wittenburg & Strawn [1].

### References

1 Peter Wittenburg and George Strawn. *Common Patterns in Revolutionary Infrastructures and Data*. 2018.
URL: http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

## 3.2 FAIR vs. Open

*Natalia Manola (University of Athens, GR)*



Open Scholarship and Open Science are becoming the modus operandi in research, but for data sharing to reach researchers a language of scholarly communication should be spoken. Excellent researchers do not have time to design infrastructure or temper technology, but

[2] http://rd-alliance.org
[3] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda/
[4] https://www.dona.net
[5] https://www.go-fair.org

they are eventually targeted towards communicating with peers through publishing any forms of research results, which contribute to their career development. FAIR and open data are now key aspects in scholarly communication, even from an early stage of research production, and our duty as infrastructure providers is to enable openness and FAIRness by design into our services and processes. Even though there is an unwarranted perception that FAIR and Open are overlapping terms, in reality "FAIR is not Open", and "Open does not imply FAIR" [1]. This adds some complexity in our communication to researchers, and we are often faced with questions of whether to aim for open or FAIR data, and in which situations would one or the other be enough? As a starting point, many of the FAIRness principles for data are prerequisites for their openness. However, data being FAIR does not directly imply that it is also open as there are "levels" of openness which are subjected to ownership, intellectual property rights, sensitivity issues, licensing etc. In practice, the FAIR principles are directed more towards technical aspects than towards moral and ethical aspects of data, especially as these address sharing by default for publicly funded research. Moreover, FAIR principles require clarity and transparency around the conditions governing access and reuse, and relevant services focus upon provisions to make data available for reuse under clearly-defined conditions and licenses, through well-defined processes, and with appropriate acknowledgement and citation. On the other hand, open does not directly mean FAIR. Open datasets without being FAIR, e.g., without proper metadata or software to access, without specifying the proper licenses are useless to their intended users. This is true for all fields, but it is more apparent in cases where we must have reproducible and accountable science, such as medical data where patient health history matters or humanities where working on already available data makes large part of the research. Furthermore, some key ethical issues still persist in openness: how do we code for machine readiness, with GDPR and data protection seen as examples; how do different domains perceive such issues and how do individuals react; where does big data and linked data come into play? In the end, costs put aside, we should aim for designing and operating infrastructures which produce "Open data in a FAIR way", considering above ethical issues.

### References

1 Mons, B. and Neylon, C. and Velterop, J. and Dumontier, M. and Da Silva Santos, L. and Wilkinson, M. 2017. *Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud*. Information Services and Use. 37 (1): pp. 49-56. DOI: 10.3233/ISU-170824

## 3.3 FAIR Principles within the EOSC

*Klaus Tochtermann (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE)*



The idea of the European Open Science Cloud (EOSC)[6] is to leverage European research data management to the next level of excellence. The EOSC will connect existing and future research data centres with one another and will offer a free point of use, open, and seamless services for storage, management, analysis, and re-use of research data. The talk highlights

[6] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

three components related to FAIR: 1) data, 2) services, and 3) infrastructure. To establish the link between FAIR and EOSC, the talk analyses relevant EOSC documents, such as the reports of the EOSC High Level Expert Groups[7] or the Implementation Roadmap for the EOSC[8], events such as the EOSC summit, and developments such as new projects like FAIRsFAIR, FAIRPlus, and EOSC-Life, which have been awarded recently with EC funding to foster the FAIR principles. Within this context the Implementation Roadmap for the EOSC recommends to develop FAIR data tools, specifications, catalogues, and standards to best support scientists and innovators, and to stimulate the demand for FAIR data through consistent FAIR data mandates and incentives to open data by research funders and institutions across Europe. With respect to infrastructures the talk addresses the need for FAIR-compliant certification schemes for FAIR data infrastructures. And finally, the talk argues for the need of initial services that are required to gather and organise FAIR data and data-related research products and which should be made accessible via a service platform.

# 4 Working Groups on Challenges of FAIR Principles for Computer Science

The participants discussed in three parallel working groups each aspect of the FAIR principles with the following questions:

- What are the current challenges?
- What are possible solutions and how can computer science support?

The following sections are brief summaries from the respective working groups.

## 4.1 Working Group 1

*Natalia Manola (University of Athens, GR / Moderator), Wilhelm Hasselbring (Universität Kiel, DE / Rapporteur), and Peter Mutschke (GESIS – Leibniz Institute for the Social Sciences – Cologne, DE / Contributor)*



### Findable

**Challenges.**

1. The data infrastructure landscape is characterised by a strong diversity and heterogeneity of data repositories and an over reliance on cataloguing. At the same time, stakeholder requirements are not really known. An overarching, one size fits all portal is missing.
2. Appropriate metadata standards and controlled vocabularies are missing as regards both common core as well as domain-specific standards.
3. An open issue is how to design identifiers and versioning (latest, history, releases), in particular how to precisely identify arbitrary subsets of data in a dynamic setting with

[7] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud-hleg
[8] https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

data being added, deleted, changed (see RDA WGDC Recommendation on Dynamic Data citation[9], with a slightly more extensive report[10].
4. Indexing is not sufficient. Field specific selections are also needed.

**Possible solutions and support of computer science.**
1. Meta search engines, supporting both multi-disciplinary and disciplinary search, and standardized search and content harvesting APIs are needed ((a) Verborgh & Dumontier (2018): A Web API Ecosystem through Feature-Based Reuse [1], (b) combination of metadata and API example: [2] from the life science community). A further help is seen in search query and metadata auto completion, e.g., enabled via machine learning in "data lakes".
2. A common representation of metadata (see EDMI from EOSCpilot[11] and schema.org) as well as intelligent assistance for metadata creation (see Ted Nelson's vision of a literary machine for science[12] is urgently required.
3. Computer science could help by providing standard components and engineering support.
4. Leveraging industry-scale markup like schema.org (e.g. bioschema.org) would alleviate this problem.

### Accessible

**Challenges.**
1. A major problem from user perspective is seen in dead links and the lack of reliable mechanisms to deal with dead links in a sustainable way.
2. A further problem is seen in the great amount of heterogeneity of authentication, authorisation and identification processes.

**Possible solutions and support of computer science.**
1. Audit trail management services are required that provide evidence and quality control of access data and links and by this a greater transparency of data accessibility t the user.
2. A standardised protocol for authentication, authorisation, and identification (AAI) is strongly needed (e.g. a data passport containing a decision tree providing a meaningful response to users on what to do to get access).

### Interoperable

**Challenges.**
1. The variety of data formats and data encoding methods (e.g. different time/date formats, time zones) as well as the lack of format validation is seen as the major problem.
2. An infrastructure to annotate data for improving interoperability is missing.
3. Dealing with different vocabularies and languages is a major obstacle as regards interoperability.
4. Ensuring portability of data as well as tools between different platforms is still a challenge.

[9] https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf
[10] http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf
[11] https://eosc-edmi.github.io
[12] https://en.wikipedia.org/wiki/Literary_Machines

**Possible solutions and support of computer science.**

1. Common standards, such as W3C / schema.org, should be reused consequently. Data transformation methods (by use of semantic technologies and common markup languages, e.g. YAML, extract transform load) as well as advanced format validation tools are strongly required.
2. Tools for semantic annotations and an ontologies lookup service as a gatekeeper are needed.
3. Ontology crosswalks (LOV=linked open vocabularies[13]) and smart ontology mapping methods could alleviate this problem.
4. Container technologies to enable portability might help.

### Reusable

**Challenges.**

1. Different data protection regulations, intellectual property rights and licensing models make reusability challenging .
2. Provenance data that capture the entire lifecycle of data, i.e. the social process of "making" data (incl non-digital interactions), is often missed.
3. Peer review of data sets and data curation is needed.

**Possible solutions and support of computer science.**

1. Computer actionable licences are required to address this problem. Identifiers and versions should be FAIR as well.
2. Services that record workflow-generated provenance metadata are strongly needed.
3. Semi-automated data curation tools (workflow based) and tools for data management plans are required.

### References

1 Ruben Verborgh and Michel Dumontier. *A Web API Ecosystem through Feature-Based Reuse*. IEEE Internet Computing, Volume 22 , Issue 3 , May./Jun. 2018. DOI: 10.1109/MIC.2018.032501515

2 Carlos Horro Marcos1, John M. Hancock, Wiktor Jurkowski, Annemarie Eckes. *Brassica Information Portal and Elixir: MIAPPE & BrAPI.*Presented at Joint 25th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 16th European Conference on Computational Biology (ECCB) 2017. URL: https://doi.org/10.7490/f1000research.1114610.1

## 4.2 Working Group 2

*Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE / Moderator) and Daniel Mietchen (University of Virginia, US / Rapporteur)*



### Findable

**Challenges.** Rich metadata that includes context and provenance information described in different languages and annotations from different views will be essential to facilitate

[13] https://lov.linkeddata.es/dataset/lov/

broad findability and finally interpretation and reuse of digital objects by researchers across communities. An extension of search to content aspects would be helpful, however, we need to have simple ways to define content patterns and an infrastructure supporting content search is required. In some analyses metadata is treated as "data", therefore it makes sense to treat metadata as separate Digital Objects of special types allowing machines to carry out suitable operations. It was argued that many metadata assertions are made at different phases, but that we do not yet have suitable means to collect these assertions to come to rich metadata. These challenges pose requirements on infrastructure development and computer science.

**Possible solutions and support of computer science.**

- An infrastructure based on the concept of Digital Objects such as worked out in RDA is urgently required. This infrastructure needs to be based on an identifier system that is available for everyone and supports the large number of stable links needed.
- Repositories and portals need to offer metadata annotation frameworks such that these extensions are kept as separate digital objects, but nevertheless can be made part of the findability context.
- An increased application of automatic workflows is required to automatically generate the rich metadata required by machines. However, this step will only be done if the domain researchers get access to easy to use and flexible workflow orchestration frameworks.

### Accessible

**Challenges.** Automatic content negotiation including variants and versions is missing and creating inefficiencies for users. The rules for dealing with dynamic metadata and data have been specified clearly within RDA and the repositories need to make clear which policies with respect to versioning they are using. The current authentication mechanisms lack the support of detailed credentials such as specific "roles" which are needed for fine graded access control. Although basic technology exists, we miss an efficiently working authorisation system for several different distributed scenarios. Also, these challenges pose requirements on infrastructure development and computer science.

**Possible solutions and support of computer science.**

- Infrastructure components such as repositories need to support content negotiation and the application of the RDA rules on dynamic data.
- The currently used authentication systems need to be extended to support the needed detail of credentials.
- Effort needs to be taken urgently to design and develop practically usable authorisation solutions for distributed data scenarios such that also machines can easily find the correct information.

### Interoperable

**Challenges.** It was stated that there are many different layers of interoperability starting with the structural and semantic specification of protocols enabling the exchange of, for example, digital objects up to the encoding of phenomena in data and metadata that are close to the research topics being studied. While the first can be specified in all detail, the others are subject of changes and dispute. In addition, semantic interpretation varies substantially with the semantic distance to the original source and the vocabularies used. It was concluded that at this level, semantic precision is an illusion. The minimum that is

expected is that everyone defines and registers the schemas and concepts being used, allowing others to interpret and refer to them. Another important topic raised was the lack of means to link digital objects with operations that are suitable for their type. This led to some conclusions about the needed infrastructures and computer science actions. Especially with respect to interoperability, a gap between computer science research and urgently needed infrastructure building was considered.These challenges are huge and conceptual support from computer science is urgently required.

**Possible solutions and support of computer science.**

- Obviously a systematic and systemic approach for registries of schemas, concepts and vocabularies is required to overcome the current fragmentation. Such a solution would support users to reuse existing best practices and to optimise the implementation of "annotation by stealth" systems.
- A systemic solution to support crosswalks between different semantic spaces and to share these with other users would increase efficiency of work.
- Also, the broad availability and use of mechanisms to link types of digital objects with operations as suggested by RDA's data type registry specifications would increase efficiency for human users and open the path towards automation.

### Reusable

**Challenges.** Detailed provenance information is required in particular to make workflow orchestration and execution possible. Yet there are no best practices how to link extensive provenance information allowing machines to find it. Improved mechanisms for provenance and containers would also improve the conditions for reproducibility as a subset of reusability and to access earlier versions of digital objects. Making licenses machine readable and actionable is a challenging topic, the way smart contracts are defined needs to be standardised. Increasing the quality of data sets is another big challenge and applying checkpoints in workflows to control quality is recommended. The review of data sets is difficult given the huge amounts of data and given that only a small percentage of data will be published officially. These challenges pose requirements on infrastructure development and computer science.

**Possible solutions and support of computer science.**

- More efforts need to be taken to have mechanisms to create and link provenance information that can be broadly used.
- Improve reproducibility of research results is an urgent requirement. Here we can refer to other Dagstuhl workshops, such as Reproducibility of Data-Oriented Experiments in e-Science[14]
- Also, in the area of machine actionable licenses more conceptual effort needs to be invested to come to a limited set of best practices.
- Improving data quality is an urgent requirement, since reuse is based on trust in the content. Ways to improve and test quality need to be worked out.

[14] https://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=16041

## 4.3 Working Group 3

*Klaus Tochtermann (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE / Moderator), Kathleen Gregory (Data Archiving and Networked Services, NL / Rapporteur), and Luiz Olavo Bonino da Silva Santos (GO FAIR – Leiden, NL / Contributor)*



### Findable

**Challenges.**

1. Searchability does not directly equate to findability, particularly for human actors, who engage in "user journeys" of browsing, linking to related content, serendipitous discovery, and direct searching. A variety of data, not just data produced through research, including open or governmental data, are also of interest.
2. Metadata creation is problematic. It is often generated by humans in unstandardized ways. Even when standards exist within a discipline, there are various interpretations and variations in how metadata fields are populated. At the same time, humans also provide a necessary level of control.
3. Sometimes a PID resolves to a metadata landing page; sometimes it redirects the user/agent directly to the data itself. This discrepancy is especially problematic for machine agents.
4. Private entities (e.g. Google DataSearch, publishers) are often in charge of discovery platforms. These entities could decide to restrict or cut off access to search platforms, which would limit the discoverability of data.

**Possible solutions and support of computer science.**

1. Semantic techniques could help both human and machine agents to locate distributed data.
2. Machine-generated metadata which is subsequently annotated by humans or semi-automated metadata generation could alleviate this problem.
3. The solution to this problem is not purely technical, but also falls within the policy realm. Policies need to be made that standardize where PIDs direct agents. Developing PID information types, metadata describing the information type where a PID points, could also be a possible solution.
4. Platforms should be open source to ensure that they remain open and accessible.

### Accessible

**Challenges.**

1. If new communication protocols are required, how can it be demonstrated that they are open, free, universally implementable and allow for authentication and authorisation when necessary?
2. How long should metadata be retained for data that are no longer available?

**Possible solutions and support of computer science.**

1. Issuing certificates to individuals/agents, creating registries of access protocols, and certifying the openness of these protocols could help to automate and streamline access.
2. These metadata could be kept for a time period defined by a timestamp; at the defined time, archivists could review the metadata and decide to retain or delete the metadata.

#### Interoperable

**Challenges.**

1. The variety of metadata standards and representation languages between and within domains makes interoperability challenging.

**Possible solutions and support of computer science.**

1. Domain specific use cases exploring interoperability could help to better understand these differences. Semantic technologies, e.g., ontology alignment and applying translations between representation languages, could help to make these data interoperable.

#### Reusable

**Challenges.**

1. Provenance information is vital for reuse. Computer-actionable provenance schemas exist, but are perhaps underused.

**Possible solutions and support of computer science.**

1. The use of schemas for provenance information should be further investigated and encouraged. The code used to generate the data should also be included with the data or added to the metadata. The reuse of algorithms, software, and standards necessary for interpretation also needs to be supported.

# 5 Working Groups on Specific Types of Challenges for FAIR

## 5.1 Legal Tech: Licenses and Software to Ensure Trust

*Marie Farge (ENS – Paris and CNRS, FR / Moderator) and Ron Dekker (CESSDA ERIC, NO / Rapporteur)*



This working group discussed the complexity with legal issues regarding the access and use of research data and how technology and software in the sense of "Legal Tech" can help.

**Definitions.**

- Data: Any identifiable object is a data.
- License: It is a subset of contracts.
- FAIR: It qualifies the processes and protocols necessary to authorise a human or a machine to access data (its content and/or its metadata).
- Legal Tech: Use of computer science methods and software to help stakeholders solve
- legal issues on data production and use, e.g., on privacy and security questions, copyright and IPR (intellectual property rights), etc.
- reproducibility issues on all publications and processes, e.g., on the European GDPR (General Data Protection Regulation) issues, etc.

**Present situation.** Before internet and electronic publishing the interaction between a data owner and a data user was between two humans (who usually knew each others), or was mediated by a third human, usually a librarian. This interaction was reciprocal and based on mutual trust, without or with very little legal control.

Today, the interaction is between a data owner and a very large number of data users. They no more know each others, while today their interaction is mediated by networks, service providers, institutions, platforms (machines and software), and humans acting as brokers. Moreover, there are increasing requirements concerning privacy, security, economic or scientific value, and idiosyncratic risk (where each case requires new negotiations, new contracts or agreements, etc.).

**Needed evolution.** Key issues are trust and compliance, and today machines together with software step in. We need frameworks and infrastructures that can do verification using software. We would like to transform licenses into computational models.

Today access to confidential data is cumbersome. There is lack of trust between owner and user: on access, on use, on applications. There are many – too many – contracts, agreements, etc. Institutions act as brokers between owners and users, but this is labour intensive and complex. If machines become brokers, this will simplify offer better performance on compliance, and enable to do post-compliance (on new papers). If machines instead of humans are allowed to access data (i. e., queries via algorithms instead of access to individual data) and process them, one will no longer need to anonymise data to comply with privacy issues. Indeed, sometimes anonymisation is not possible; moreover, most of the time it induces loss of information, and it is not necessarily a guarantee for compliance.

One also needs to simplify the types or numbers of contracts/agreements and the legal and soft conditions for approval to use (e.g., on confidential data), to meet pre-conditions, post-approval compliance, etc. To achieve this, one could replace the current broker, doing checks manually, by machines and software. They would provide even more than a broker could ever do, hence provide more trust into the system.

We need to engage computer science with legal, economic, and technical communities. We need contract-language that is machine-parsable on transactions. For this one should specify the rules for audit and use distributed machine-learning. Digital Objects should also have the license-metadata with them.

Regarding the use of legal tech for reproducibility issues we left with the following open questions:

- Is there a method for reproducibility checking on confidential data done by a machine?
- Is it possible to do all reproducibility checks by a machine?

**Use cases.**

- NHS Research Passport[15]
- Image Processing Online (IPOL)[16]: Open Access publishing platform which offers the possibility to test one's own data on a given algorithm (implemented in open source) to see if it is useful for such data. Each article contains a text, an algorithm, and its source code, with an online demonstration facility and an archive of experiments. Text and source code are peer-reviewed and the demonstration is controlled. Moreover, the software of the publishing platform is open source and available on GitHub.

[15] https://www.nihr.ac.uk/about-us/CCF/policy-and-standards/research-passports.htm
[16] https://www.ipol.im

## 5.2 (Self-)Improvement of FAIRification

*Carole Goble (University of Manchester, GB / Moderator & Rapporteur) and Michel Dumontier (Maastricht University, NL / Contributor)*



The working group addressed the question: How "FAIRification" can be usefully assessed? A central focus of the discussion was about concerns in FAIR assessments and certifications.

The FAIR principles are aspirational – they articulate a set of desirable properties in digital objects to increase their potential to be discovered and reused by others. Achieving the vision of an Internet of FAIR digital objects will pose a substantial challenge to create it in a sustainable manner. Some aspects of the FAIR principles are readily achievable, while others may entail substantial and sustained effort. The likelihood that any given resource will completely fulfil the FAIR principles out of the box or at any given time is low – but that is to be expected and not a negative situation, and offers the opportunity to improve its value to a wider community of (re)users. Therefore, the role of any assessment tool should not be to "judge" a repository, but to provide indications of what can be expected from a resource.

Towards obtaining a picture of the state of FAIRness in digital objects, initiatives so far range from the development of questionnaires to elicit self reflection to gather largely qualitative assessments to metric-based software aim to gather evidence of quantitative adherence. Indications of what can be assessed as span this quantitative/qualitative spectrum. The language ranges from "metrics", understood as numerical (or ordinal) measures of quantitative assessment for comparison and compliance, to "indicators" as an attempt to embrace a range of signals beyond those that can be readily counted, and to incorporate non-mechanistic means of assessment that take into account the costs and return on investment of FAIRification of datasets by data providers. The working group highlighted the challenges of communicating FAIR by overly simplistic methods such as star ratings, as exemplified in early experiments with the DANS FAIR Assessment tool[17], for example.

Others, such as GO-FAIR, are examining the feasibility of FAIR certification via nationally accredited third-parties that could apply to datasets, repositories, software, services (such as training), organisations, and people (such as FAIR data stewards). Certification involves the confirmation of certain characteristics of an object, person, or organisation. Certifications aim to establish trust, set expectations in terms of quality and utility, offer choice, encourage criticism and roadmaps for improvement. An example of certification is The Core Trust Seal (CTS), which offers certification for online repositories, highlighted by the European Commission's High Level Expert Group on Turning FAIR into Reality. The working group cautioned against premature certification against assessment criteria that has yet to be fully understood or have buy-in from communities. Dangers highlighted included: favouring one community over another (for example digital librarianship and funder compliance over domain specialist data providers) and one actor over another (for example dataset consumers over dataset providers).

While well intentioned, current FAIR assessments and certification schemes only provide a narrow picture of features that may be needed to fully realize the FAIR vision. At NETTAB 2018, Christine Durinx, Executive Director SIB & co-lead of ELIXIR Data Platform, presented

[17] https://www.surveymonkey.com/r/fairdat

an expanded set of indicators that better captures the goals of provisioning of high quality database service (http://www.igst.it/nettab/2018/files/2018/10/NETTAB2018_Durinx.pdf). This included the use of persistent and unique identifiers, number and growth of entries in the repository, technical performance of system, use of community-recognised standards for (meta)data, documentation of provenance, availability of data, and customer service. In this way, the FAIR principles are an important, but underspecified set of requirements. Moreover, there is serious concern that a resource that does not adhere to the FAIR assessment or certification could be seen as having lower quality or value than a resource that ticks all the boxes. Additionally, fully meeting the FAIR principles may be prohibitively expensive for individuals and particular organisations. This is particularly of concern for organisations that are under pressure from funding agencies to be FAIR, while other well justified concerns prioritise their efforts.

Another promising option is to consider FAIR as a contract. By this we mean that FAIR is used as a mechanism for reporting, for expectation management of consumers, and as a roadmap. Contracts in every aspect of FAIR would enable an agent to reliably assess what can be expected from a resource and to decide if it should use the resource. Roadmaps and frameworks are helpful for systematic reviews. Review helps decision making and direction setting. For these reasons, the working group was in favour of developing "FAIRification Roadmaps" that would foster positive discussions with stakeholders on improving FAIRness with a sensitivity to the value-returning activities. The return on investment made by all stakeholders (repository owners, data consumers, funding agencies) is a key part of a "FAIRification maturity model". Established computer science principles can help to implement corresponding models, e.g., based on CMMI[18] and related assessments. See also the FAIR Capability Maturation Model which is developed in the FAIRplus project[19] as example.

The debate regarding FAIR assessment highlighted the spectrum of agendas, viewpoints, and contexts that need to be taken into account:

- from data consumers to data providers
- from objective automated FAIR assessments to manually mediated assessments incorporating subjectivity
- from focus on supporting automated analytics to supporting human interaction and decision making
- from FAIR assessments confined to technical aspects to those that incorporate social, political, and economic aspects (particularly for data providers)
- from a set of rules and tick lists to a roadmap or contract

**Examples and use cases.**

- FAIRmetrics[20]
- FAIRshake tool[21]
- ANDS FAIR Data Self-Assessment Tool[22]
- CMMI[23]
- DANS FAIR Assessment Tools[24]

[18] https://en.wikipedia.org/wiki/Capability_Maturity_Model_Integration
[19] https://fairplus-project.eu
[20] http://fairmetrics.org
[21] http://fairshake.cloud
[22] https://www.ands.org.au/working-with-data/fairdata/fair-data-self-assessment-tool
[23] https://en.wikipedia.org/wiki/Capability_Maturity_Model_Integration
[24] https://dans.knaw.nl/en/projects/projects

- Science Europe Guidelines[25]
- ELIXIR CDR criteria[26]
- Core Trust Seal[27]

### References


1 Jerry Z. Muller. *The Tyranny of Metrics*. Princeton, 220 pp, February 2018. ISBN: 978 0 691 17495 2

2 James Wilsdon et al. *The Metric Tide Sage*. 168 pp, February 2016. ISBN: 978 1 4739 7306 0

3 Christine Durinx, Jo McEntyre, Ron Appel, Rolf Apweiler, Mary Barlow, Niklas Blomberg, Chuck Cook, Elisabeth Gasteiger, Jee-Hyub Kim, Rodrigo Lopez, Nicole Redaschi, Heinz Stockinger, Daniel Teixeira, Alfonso Valencia. *Identifying ELIXIR Core Data Resources*. URL: https://f1000research.com/articles/5-2422/v2

4 Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos & Michel Dumontier. *A design framework and exemplar metrics for FAIRness*. URL: https://www.nature.com/articles/sdata2018118

5 *Turning FAIR into reality*. Final report and action plan from the European Commission expert group on FAIR data. Nov 2018. ISBN: 978-92-79-96546-3 DOI: 10.2777/1524

6 Alastair Dunning, Madeleine de Smaele, Jasmin Böhme. *Are the FAIR Data Principles Fair?* URL: http://www.ijdc.net/article/view/567


## 5.3 FAIR Principles & Data Intensive Science

*Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE / Moderator) and Kees den Heijer (TU Delft, NL / Rapporteur)*



The big question that was addressed in this working group was whether data intensive science (DIS) poses different requirements with respect to the FAIR principles as other types of data usage. Since definitions of terms such as DIS and Big Data are fuzzy, the group, consisting of experts working on large data sets and computer scientists, first briefly characterised what is meant by DIS. Attributes were mentioned such as the 4 Vs (Volume, Variety, Velocity, Veracity), large data sets with numerical data generated by sensors and simulations, an increased obligation to use automatic workflows which need to be sufficiently flexible, an area of numerical transformation to integrate data from different sources, application of complex statistical AI methods to first extract knowledge before it can be exploited using semantic technologies, and existence of the iceberg phenomenon where much data is being created and reused from collaborators far before subset collections will be published.

In different scenarios, the terms used in the FAIR principles need to be interpreted in the respective contexts:

- Finding suitable data sets, for example, for training models by machine learning requires very rich metadata including provenance information. If search would be extended to

[25] https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf
[26] https://www.elixir-europe.org/platforms/data/core-data-resources
[27] https://www.coretrustseal.org

content search, a broadly available HPC infrastructure would be required which does not exist.
- The need for increased richness of metadata leads to more complex search interfaces which researchers in general do not like, i.e., interface engineering to camouflage complexity is not trivial.
- The interpretation of the term "interoperability" as introduced in the FAIR principles needs[28] to be amended since the integration of typical data in DIS from different sources mostly requires extensive numerical transformations and normalisations due to differences between the underlying models and the calculations of missing data. At that stage one cannot speak of the application of "knowledge representation languages", instead we speak about structures and formats with headers informing about the variables.
- In DIS there is an urgent need to turn to automatic workflows with repeating patterns and extensions to create the necessary documentation. These, however, need to be flexible and parameterised, and to allow accessing subsets of data, demand specific knowledge to be created which is beyond the knowledge of the domain specialists.
- A reliable and highly performant system to register and resolve PIDs that will be available for many decades, that will be independent of protocol specifications which may change over time and that is free of misleading semantics such as included in URLs.
- Professional mechanisms to allow code to be moved to data and to be executed remotely are a must posing complex organisational and administrative challenges.
- The availability of smart VREs supporting the various functions would reduce the complexity of the tasks for the researchers.

Summarising, it was agreed that DIS requires specific interpretations of a few terms used in the FAIR principles and are stressing the needed infrastructures and services due to the scale of data and calculations. Computer science would have to collaborate closely with domain scientists to work out solutions for the various challenges indicated.

# 6 Working Groups on Identified Topics for the Dagstuhl Manifesto

## 6.1 Infrastructures & Services Aspects

*Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE / Moderator) and Dieter Van Uytvanck (CLARIN ERIC – Utrecht, NL / Rapporteur)*



This working group went through all notes to indicate relevant topics for building and extending infrastructures and the service landscape. Although building large infrastructures includes politics, socio-economics, and technology, the discussions focused mainly on the latter. Many points were addressed in the different sessions, here we can only mention a few:
- The need for a PID registration and resolution infrastructure available for every researcher (and beyond) and fulfilling several requirements (powerful, protocol-independence, stability of references over decades, support of standardised attributes, support of binding of digital object components to facilitate interoperability, inclusion of passport information, etc.).

[28] I1. (meta)data use a *formal, accessible, shared, and broadly applicable language* for knowledge representation

A PID system needs to be based on interoperability standards such as ITU X.1255 and a standardisation of attributes as done in RDA.
- A set of ready-made services to continuously monitor the availability and state of URI and PIDs and make use of its possibilities would be extremely helpful (link checking, cross-referencing, integrity checking, etc.). Since these checks can be resource-intensive (e.g. computing, network) some degree of centralization, as to prevent duplication, can provide a higher efficiency.
- It is important to have improved support for rich metadata creation, exposure, searching, and mapping, including provenance facilitated by registered schemas and semantics. A limited set of serialisation formats should be accepted, such as XML, JSON, RDF, etc.
- Since the creation of metadata is a demanding task, automatic extraction of metadata on the basis of data files should be considered. A landscape analysis of such available technologies would be valuable.
- We need a systematic and systemic approach to schema, concept, and vocabulary registration allowing people to easily register, find, and reuse them. RDA offers such a schema registry. This can be seen as a basis for much better support of semantic crosswalks.
- Changes to data and metadata must be versioned and traceable to understand the state of the data at their (various) time(s) of use.
- A systematic solution allowing registering types of digital objects and link sets of operations with it, such as developed in RDA, is highly needed to foster automation.
- A much more improved authentication and authorization infrastructure addressing different distributed scenarios is highly required. This must be amended by computer-readable license consent (smart contracts).
- Much more support for creating flexible workflows is required. These should also facilitate reproducibility and should make use of state of the art packaging formats, as suggested, for example, by Research Objects[29].
- A move towards self-explanatory APIs that include semantic annotations is needed to facilitate machine action. They are digital objects having a PID and descriptions.
- Much better support for the scenario where code is transferred to the data to be executed is required.
- Finally it was agreed that suitable VREs bundling the access to infrastructures and its many services would offer great opportunities for more easy access to all features by researchers.

## 6.2 Computer Science Research Topics

*Achim Streit (KIT – Karlsruher Institut für Technologie, DE / Moderator) and Tobias Weigel (DKRZ Hamburg, DE / Rapporteur)*



This working group discussed research topics in computer science related to implement FAIR data infrastructures. This covers a broad spectrum of subfields from mathematical foundations, algorithms and data structures, security, AI, software engineering, applied

[29] http://www.researchobject.org

computer science up to theory of computation (cf. Wikipedia page on computer science[30]). The working group put a focus of its discussion in defining future computer science research topics in the form of PhD topics. These were:

- Tamper-proof exchange and tracking of digital objects across distributed infrastructures
- Provenance capturing and reasoning on provenance data to enable automated data identification and integration across heterogeneous domains
- Privacy-preserving analytics across distributed data sets
- Demonstration of a fully automated closed loop research cycle system, from hypothesis generation, data identification, analysis, hypothesis verification to novel hypothesis derivation
- Automated informed consent negotiation and inference
- Intelligent content-based searching using AI and performance benchmarking with state-of-the-art metadata-powered search techniques
- Visual analytics in support of data finding – new forms of Human-Computer-Interaction based on interdisciplinary research with Arts and Social Sciences
- Quantifiable FAIR-ability of architectural frameworks of data infrastructures
- Representation of ethics and moral in technical solutions in FAIR data infrastructures
- Semantic* in support of intelligent FAIR services and based on ontologies and vocabulary crosswalking
- Security & Privacy frameworks for making data FAIR
- Framework to analyse the impact of FAIR metrics
- Novel data management/storage concepts enabling persistent provisioning of large-scale research data across evolving versions aggregated over long time scales
- New approaches for digital preservation to ensure FAIR data remains FAIR over long periods of time
- Social Machines and FAIR: Crowdsourcing FAIR
- Trust and identity in the context of FAIR data infrastructures

## 6.3 FAIR Computer Science Research

*Wilhelm Hasselbring (Universität Kiel, DE / Moderator & Rapporteur) and Paolo Manghi (ISTI-CNR – Pisa, IT / Contributor)*



After focussing on how computer science can contribute to implement the FAIR principles in several working groups, this working group addressed the question how are the FAIR principles are currently adopted by computer science itself. The group started to discuss examples from software engineering.

In software engineering everything you can store as file is called an artifact. Most major software engineering conferences meanwhile offer artifact evaluations for papers accepted to the conference main programme. ACM SIGMOD, for instance, calls its artifact evaluation as reproducibility evaluation. Papers that also have artifact evaluation have more citations [1]. In some cases, computer science experiments are not reproducible and only repeatable (e.g. performance research, you need the same hardware to reproduce results). See also [2].

[30] https://en.wikipedia.org/wiki/Outline_of_computer_science

Based on this input the group started to collect aspects and issues to make computer science research FAIR.

**Findable.**

- Publish as much as possible. Artifacts are: software, data, employed methods, workflows, protocols, services, virtual machines/containers, documents, etc.
- Assign PIDs to everything (not necessarily DOIs).
- (Pre-)Publication of scientific processes/workflows (e.g. protocols.io).
- Use repositories like DockerHub.
- Software metadata remains a great challenge: software citation (SSI, OpenAIRE, CodeMeta), software identification (RDA Group).
- Executable papers and echanced publications [3].

**Accessible.**

- Research artifacts should be published with preservation in mind. GitHub, for example, does not do that. Publishing involves citation and preservation (e.g. Zenodo.org).

**Interoperable**

- Portable software is required (subproblem of software preservation).

**Reusable.**

- Use (and describe in the metadata) standard tools for sharing scientific thinking. Containers (Docker etc) and virtualisation help. Same for Jupyter notebooks. Use workflow languages such as the common workflow language (CWL)[31].
- Allowing repeatability by offering cloud-based services, such as VREs for example. Distinguish between Software-as-Code (e.g., via GitHub) and Software-as-a-Service (e.g., via some cloud service).
- Follow standards for APIs and metadata. Documentation is essential.
- Conference artifact evaluation processes already help to check quality via peer review.

**References**

1 Bruce R. Childers; Panos K. Chrysanthis.*Artifact Evaluation: Is It a Real Incentive?*. 2017 IEEE 13th International Conference on e-Science (e-Science). DOI: 10.1109/eScience.2017.79

2 Shriram Krishnamurthi and Jan Vitek. *The real software crisis: repeatability as a core value*. Commun. ACM 58, 3 (February 2015), 34-36. DOI: https://doi.org/10.1145/2658987

3 Bardi, A. and Manghi, P. *Enhanced Publications: Data Models and Information Systems*. LIBER Quarterly, 23(4), pp.240–273. DOI: http://doi.org/10.18352/lq.8445

## 7 Acknowledgements

We thank all participants of the workshop for their valuable contributions and the Dagstuhl team who have made this a successful event.

[31] https://www.commonwl.org

## Participants

- Marcel R. Ackermann
LZI Schloss Dagstuhl & dblp
Trier, DE
- Luiz Dlavo Bonino da Silva Santos
GO FAIR – Leiden, NL
- Timothy W. Clark
University of Virginia, US
- Ron Dekker
CESSDA ERIC, NO
- Kees den Heijer
TU Delft, NL
- Michel Dumontier
Maastricht University, NL
- Marie Farge
ENS – Paris and CNRS, FR
- Sascha Friesike
VU University of Amsterdam, NL
- Carole Goble
University of Manchester, GB
- Kathleen Gregory
NL
- Gregor Hagedorn
Museum für Naturkunde – Berlin, DE
- Wilhelm Hasselbring
Universität Kiel, DE
- Oliver Kohlbacher
Universität Tübingen, DE
- Paolo Manghi
ISTI-CNR – Pisa, IT
- Natalia Manola
University of Athens, GR
- Daniel Mietchen
University of Virginia, US
- Peter Mutschke
GESIS – Leibniz Institute for the Social Sciences – Cologne, DE
- Heike Neuroth
FH Potsdam, DE
- Andreas Rauber
TU Wien, AT
- Marc Rittberger
DIPF – Frankfurt am Main, DE
- Raphael Ritz
Max Planck Computing and Data Facility – Garching, DE
- Guido Scherp
ZBW-Leibniz-Informationszentrum Wirtschaft – Kiel, DE
- Birgit Schmidt
SuB – Göttingen, DE
- Achim Streit
KIT – Karlsruher Institut für Technologie, DE
- Klaus Tochtermann
ZBW-Leibniz-Informationszentrum Wirtschaft – Kiel, DE
- Dieter Van Uytvanck
CLARIN ERIC – Utrecht, NL
- Tobias Weigel
DKRZ Hamburg, DE
- Mark D. Wilkinson
Polytechnic University of Madrid, ES
- Peter Wittenburg
Max Planck Computing and Data Facility – Garching, DE

Report from Dagstuhl Seminar 18481

# High Throughput Connectomics

**Edited by**

**Moritz Helmstaedter[1], Jeff Lichtman[2], and Nir Shavit[3]**

1 **MPI for Brain Research – Frankfurt am Main, DE,** moritz.helmstaedter@brain.mpg.de
2 **Harvard University – Cambridge, US,** jeff@mcb.harvard.edu
3 **MIT – Cambridge, US,** shanir@csail.mit.edu

## Abstract

The structure of the nervous system is extraordinarily complicated because individual neurons are interconnected to hundreds or even thousands of other cells in networks that can extend over large volumes. Mapping such networks at the level of synaptic connections, a field called connectomics, began in the 1970s and has recently garnered general interest thanks to technical and computational advances that offer the possibility of mapping mammalian brains. However, modern connectomics produces 'big data' that must be analyzed at unprecedented rates, and will require, as with genomics at the time, breakthrough algorithmic and computational solutions. This workshop will bring together key researchers in the field, and experts from related fields, in order to understand the problems at hand and provide new approaches towards the design of high throughput systems for mapping the micro-connectivity of the brain.

**Seminar** November 25–30, 2018 – http://www.dagstuhl.de/18481
**2012 ACM Subject Classification** Computing methodologies → 3D imaging, Applied computing → Biological networks, Computing methodologies → Image processing, Computing methodologies → Image segmentation
**Keywords and phrases** Big Data, Connectomics, Distributed Computing, Machine Learning, Parallel Computing
**Digital Object Identifier** 10.4230/DagRep.8.11.112
**Edited in cooperation with** Daniel R. Berger

## 1 Executive Summary

*Nir Shavit (MIT – Cambridge, US)*



Our workshop brought together experts in the computational aspects of connectomics. A week of lectures and work-group meetings in a lively and collegial environment led to a collection of interesting conclusions. One big idea that was put forth in the meeting was the gargantuan effort of reconstructing a complete mouse brain. Another was to completely map the white matter connectivity of a mammalian brain. We also discussed which techniques/pipelines we should continue to pursue as a community. In that vein one big conclusion was that you have to have both the engineers and software working on a pipeline; distributing software only is not sufficient (you need dedicated engineers to run the software, it can't be based just on grad students). Zeiss reported on a multibeam 331 beam microscope that was in the making. There were also discussions on quality measures and metrics for connectomics reconstruction, and on developing standardized datasets for segmentation training and comparison of algorithms

(scaling up from current day small datasets). Finally, there were discussions on the ethics and policies in the area going forward – Should we rely more on industrial partners to provide compute power and storage, or is it better to keep most of the research in universities and non-for-profit research institutes.

## Introduction

The sheer complexity of the brain means that sooner or later the data describing brains must transition from something that is rather easily managed to something far less tractable. This transition appears to now be underway. The accumulation of ever-bigger brain data is a byproduct of the development of a number of new technologies that provide digitized information about the structural organization (anatomy) and the function of neural tissue. These new collection approaches bring novel data into neuroscience that potentially bears on many poorly understood aspects of the nervous system. Fundamental scientific questions such as the way learned information is instantiated in the brain and how brains change over the course of development and aging are likely to be usefully addressed in the coming decades as large data sets mapping networks of neurons at high resolution become available.

Mapping networks of neurons at the level of synaptic connections, a field called connectomics, began in the 1970s with a the study of the small nervous system of a worm and has recently garnered general interest thanks to technical and computational advances that automate the collection of electron-microscopy data and offer the possibility of mapping even large mammalian brains. However, modern connectomics produces 'big data', unprecedented quantities of digital information at unprecedented rates, and will require, as with genomics at the time, breakthrough algorithmic and computational solutions.

Unfortunately the generation of large data sets is actually the easy part. Our experience in the nascent field of connectomics indicates that there are many challenges associated with the steps after data acquisition, that is, the process of turning the data into a mineable commodity. This workshop will focus on addressing these challenges by bringing together researchers developing algorithms and deploying software systems that enable high-throughput analysis of connectomic data.

While high-throughput connectomics must tackle many of the problems that occur in big data science and engineering, tremendous differences in data size, computational complexity, and the problem domain will require novel computational frameworks, algorithms, and systems. Input image data in connectomics is reaching, even in its initial stages, petabytes in size at a terabytes-per-hour rate, and currently requires millions of cycles of computation per pixel. Such data is not easily moved or stored, and so on-the-fly analysis of the data as it comes off the microscope is the most likely future solution. Achieving the kind of throughput that will allow us to process the data at the rate at which it is being generated necessitates a three orders of magnitude reduction in cycles per pixel, compared to the status quo. Furthermore, there is locality to the data. Unlike other big data problems, which can often be represented as independent key-value pairs spread across many machines, reconstruction of neural circuits requires frequent data exchanges across adjacent image regions. Buffering all the data in machine memory is infeasible, as is data replication on multiple servers. That means one cannot rely on Moore's law and parallelism across data centers to solve this problem–we need to be smarter.

In a nutshell, a connectomics data set is a collection of images taken on a volume of brain tissue that has been sectioned into many thousands of small slices, each only a few

tens of nanometers thick. These slices are then imaged using custom electron microscopes to produce an image stack that will in the near future reach petabytes in size. Using one of the standard electron microscopy pipeline approaches, the key computational problems that must be addressed in order to turn the raw acquired digitized images into a useful form of "onnectivity graph" are stitching, alignment, neuron reconstruction, and synapse detection. Each digitized image tile needs to be stitched together with neighboring tiles to form a composite image of a slice. Then, the stitched slice image is aligned with the previous and subsequent slice images. Despite being mostly similar, image alignment is challenging because typically a conveyor belt collects the slices and each may rotate a few degrees, or stretch depending on its thickness. Fortunately, because of the high image resolution, alignment is practical, as axons and dendrites are readily visible in cross-section and can be traced from one section to the next. A second challenge is that, once the image data is aligned, the sectioned objects must be individuated. In these data sets, the objects are neurons and other cellular entities that are interwoven in the three-dimensional space of the sample tissue. The reconstruction of neural processes as they pass from one section to the next is directly related to the computer vision problem of obtaining a segmentation of an image series, that is, the labeling of pixels in the images according to which cell they belong to.

Although considerable progress has been achieved in computer-based image segmentation in the last few years, reliable automatic image segmentation is still an open problem. Automating the segmentation of connectomic data is challenging because the shapes of neural objects are irregular, branching, non-repeating and intertwined. Moreover, the actual number of different objects and their synaptic interconnections in a volume of brain tissue is unknown and, at the moment, even difficult to estimate or bound. Segmentation of a standard electron microscopy image is further complicated by the fact that the range of pixel intensity values of cell membranes overlaps with that of other organelles. Thus, simple thresholding to find cell boundaries does not work.

In the eyes of many, the term big data is synonymous with the storage and analysis of massive collections of digital information. The "big" refers to the size of the input sets, typically ranging in the tens or even hundreds of terabytes, and arriving at rates of several tens or hundreds of gigabytes per second. In connectomics, the size of the input set is at the high end of the big data range, and possibly among the largest data ever acquired. Images at several nanometers resolution are needed to accurately reconstruct the very fine axons, dendrites, and synaptic connections. At this resolution, a cubic mm is about 2 petabytes of data. A complete rat cortex including some white matter might require 500 cubic mm and thus would produce about an exabyte (1000 petabytes) of data. This amount is far beyond the scope of storage that can be handled by any system today (as a reference point, consider that Walmart or Aldi's database systems manage a few petabytes of data). A complete human cortex, 1000-times that of a rodent, will require a zetabyte (1000 exabytes) of data, an amount of data approaching that of all the information recorded in the world today. Obviously this means that the goal of connectomics will not be to acquire complete human brains and that for the near future one must consider reconstructions of neuronal substructures as opposed to whole brains. Moreover, it is clear that as we go beyond a few millimeters, one cannot store the raw data: it must be analyzed on the fly as it comes off the microscope and then discarded, keeping the physical tissue sample for re-imaging if needed.

What is this on-the-fly acquisition rate? The new multi-beam electron microscopes currently produced by Carl Zeiss LLC have a staggering throughput approaching 400 sections per day or a terabyte of data per hour, placing them at the far end of the big data rate spectrum. This rate, if it can be matched with appropriate reconstruction algorithms, will

allow researchers to process a cubic mm of rodent brain, that is, 2 petabytes of data, in about 6 months operating 24 hours a day, 7 days a week. Whatever computational pipeline is used to extract the connectomics graph from the image data, it will eventually have to work on the fly, at the pace of the microscope that generates this data.

The algorithms and computational techniques for developing such high throughput connectomics pipelines are the target of this workshop. The massive amounts of storage and computation require expertise not only in computational neurobiology, machine learning, and alignment techniques, but also in parallel computation, distributed systems, and storage systems. There are several groups of researchers around the world that specialize in collecting the electron microscopy datasets, and several that engage in developing matching computational pipelines. Our aim is to bring these researchers together for an extended 5-day brainstorming session. We will also invite some top researchers in related fields such as machine learning, computer vision, distributed systems, and parallel computing. Our goal for this meeting is to both build an understanding of the state of the art in high-throughput connectomics pipelines, and to brainstorm on how to move the field forward so that high throughput connectomics systems become widely available to neurobiology labs around the world.

Concretely, we would like to come out of this workshop with a hierarchical plan for future connectomics systems that solve existing systems' problems. We will begin the workshop by having workgroups discuss these problems in existing systems and then dedicate the latter part to collectively working out solutions. We will consider three levels:

1. The system layer: how data is stored, moved around and computed on in a distributed and parallel fashion.
2. The pipeline layer: how processing progresses from stitching through alignment and reconstruction.
3. The algorithm layer: the specific machine learning and error detection and correction algorithms used in various pipeline stages to bring the datasets to analyzable connectivity graphs.

Our plan is to discuss each of these in detail, with the hope of concluding the workshop with a coherent plan on how to proceed.

## Relation to previous Dagstuhl seminars

To the best of our knowledge there have been no similar Dagstuhl seminars in the past. The field of connectomics is a young cutting edge big data research area that will have important implications on both computation in the sciences (and in particular on the use of large scale machine learning in the sciences) and on artificial intelligence (through the development of new neural network models based on the neurobiological discoveries this research may lead to). We believe it is important for modern computer science to engage in such interdisciplinary applications of computing and algorithms and we are therefore eager to initiate this new seminar direction.

## 2 Table of Contents

**Working groups**

# 3 Overview of Talks

## 3.1 VAST – Efficient Manual and Semi-Automatic Labeling of Large 3D Image Stacks

*Daniel R. Berger (Harvard University – Cambridge, US)*



Recent developments in serial-section electron microscopy allow the efficient generation of very large image datasets but analyzing such data poses challenges for software tools. Here we introduce VAST (Volume Annotation and Segmentation Tool) [1], a freely available utility program for generating and editing annotations and segmentations of large volumetric image (voxel) data sets. It provides a simple yet powerful user interface for real-time exploration and analysis of large data sets even in the Petabyte range.

### References

1 Berger D. R., Seung, H. S., & Lichtman, J. W. (2018). *VAST (Volume Annotation and Segmentation Tool): efficient manual and semi-automatic labeling of large 3D image stacks.* Front Neural Circuits. 2018; 12: 88.

## 3.2 A Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster

*Davi Bock (Howard Hughes Medical Institute – Ashburn, US)*



In this talk Davi Bock described a new electron microscopic dataset of the complete adult fruit fly brain [1], and collaborative efforts to reconstruct the circuits in it.

### References

1 Zheng Z., Lauritzen J. S., Perlman E., Robinson, C. G., Nichols M., Milkie D., Torrens O., Price J., Fisher C. B., Sharifi N., Calle-Schuler S. A., Kmecova L., Iqbal J.Ali I. J., Karsh B., Trautman E. T., Bogovic J. A., Hanslovsky P., Jefferis G. S. X. E., and Bock D. D. (2018). *Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster*. Cell, Volume 174, Issue 3, 26 July 2018, Pages 730–743.e22

## 3.3 Drosophila larva EM

*Albert Cardona (University of Cambridge, GB)*



Albert talked about the analysis of circuitry in his drosophila larva EM dataset. [1]

**References**

1 Gerhard, S., Andrade, I., Fetter, R. D., Cardona, A., and Schneider-Mizell, C. M., (2017). *Conserved neural circuit structure across Drosophila larval development revealed by comparative connectomics*. eLife 2017;6:e29089.

## 3.4 Analysis Infrastructure

*Forrest Collman (Allen Institute for Brain Science – Seattle, US)*



This talk described the data analysis pipeline of the large-scale 3D electron microscopy study of mouse cortex currently done at the Allen Brain Institute in Seattle. This includes alignment, automatic segmentation, and proofreading.

## 3.5 Towards a neuronal wiring diagram of a cubic millimeter of mouse cortex

*Nuno Maçarico da Costa (Allen Institute for Brain Science – Seattle, US)*



This talk described an ongoing project at the Allen Institute for Brain Research in Seattle, which aims at reconstructing the neuronal wiring diagram of a cubic millimeter of mouse cortex from serial-section transmission electron microscopy data. Nuno also presented first results of reconstructed basket and chandelier neurons.

## 3.6 Beyond Connectomics

*Winfried Denk (MPI für Neurobiologie – Martinsried, DE)*



This talk addressed the possibility to understand the wiring diagram of the brain by unraveling how genes drive proteins, which in turn drive brain development and neuronal wiring. For this it would be very helpful to know where each protein is in the nervous system, over time. This can to some extent be approached with electron microscopy. Winfried showed that the location and orientation of individual protein structures (for example, ribosomes) can be identified in Cryo-EM image volumes by template matching [1].

**References**

1 Rickgauer J.P., Grigorieff N., and Denk, W (2017). *Single-protein detection in crowded molecular environments in cryo-EM images*. eLife 2017;6:e25648 DOI: 10.7554/eLife.25648

## 3.7 Multi-modal parcellation using mesoscale connections

*Eva Dyer (Georgia Institute of Technology – Atlanta, US)*



This talk discussed methods to analyze high-resolution 3D x-ray datasets (micro-CT) of brain sections to find cell bodies, blood vessels and more.

## 3.8 Connectivity determines neural computations in the olfactory bulb

*Rainer W. Friedrich (FMI – Basel, CH), Christel Genoud (FMI – Basel, CH), and Adrian Wanner (Princeton University, US)*



We measured odor-evoked activity in the olfactory bulb (OB) of a zebrafish larva by multi-photon calcium imaging and subsequently reconstructed all 1047 neurons and their synaptic connections by serial block face scanning electron microscopy (SBEM) and manual annotation. This comprehensive dataset allowed us to assess the contribution of neuronal connectivity to transformations of distributed neuronal activity patterns. The OB receives sensory input from the nose via an array of discrete input channels, the olfactory glomeruli. Each glomerulus is activated by a specific spectrum of ligands, and each odor is represented by a specific pattern of activation across the glomerular array. However, these combinatorial odor representations are suboptimal because chemically similar odorants evoke highly overlapping activity patterns that are difficult to distinguish by a simple classifier. Neuronal circuits within the OB decorrelate such overlapping odor representations and thereby facilitate odor classification. Previous work showed that this pattern decorrelation requires specific network connectivity that is likely to depend on the tuning curves of individual neurons. We found that the larval olfactory bulb contains a "core circuitry" that corresponds to the superficial interneuron network in the adult olfactory bulb and shows similarities to the insect antennal lobe. Long-range inter-glomerular projections are not random but organized by the identity of olfactory glomeruli. Interneurons connect ensembles of neurons responding to odorants with specific physico-chemical features through specific connectivity motifs. Analyses using matrix algebra suggested that this connectivity can, at least in part, account for the pattern decorrelation observed in the OB. Consistent with this conclusion, the connectivity reproduced experimentally observed dynamics of neuronal activity and the associated computations when implemented in a simple biophysical network model. Further analyses of the model demonstrated that the observed connectivity motifs were essential for pattern decorrelation and other computations. The wiring diagram therefore contains specific structure that removes predictable correlations from odor-evoked input patterns and supports additional computations. Hence, our "dense functional connectomics" approach revealed multisynaptic connectivity motifs in the OB that are computationally relevant and difficult to analyze by other approaches. This connectivity mediates fundamental neuronal computations that support the classification of odor-evoked activity patterns.

## 3.9 Circuit reconstruction in flies

*Jan Funke (Howard Hughes Medical Institute – Ashburn, US) and Julia Buhmann (Universität Zürich, CH)*



We present an automated method for the identification of synaptic partners in insect brains. The main idea of the method published in [1] is the direct identification of voxels that are pre-and postsynaptic to each other using a 3D U-Net. With an extension of the model of [1], we present preliminary results on a larger cutout of the FAFB dataset. We show that our method is scalable and produces qualitatively promising results.

### References

1 Buhmann, J., Krause, R., Lentini, R. C., Eckstein, N., Cook, M., Turaga, S., & Funke, J. (2018). *Synaptic partner prediction from point annotations in insect brains.* MICCAI

## 3.10 EM pipeline at FMI

*Stephan Gerhard (FMI – Basel, CH)*



A connectomics pipeline consists of a number of stages, including sample preparation, data acquisition, alignment, segmentation, proofreading, circuit analysis & visualization and data dissemination. Often, these stages are conceived as a feedforward process and feedback pathways are not often considered. I introduced the Volume Image Environment VIME (github.com/vime, unpublished), a Python-based, client-server framework for data management and linear and non-linear alignment procedures. Importantly, VIME can be interfaced with SBEMimage, a novel data acquisition software for SBEM datasets [1] enabling on-the-fly alignment and a feedback path to control acquisition based on quality control metrics from alignment. I introduced the results of processing an adult zebrafish olfactory bulb dataset with alignment in VIME, and with Google's Flood-Filling-Networks subsequently for segmentation. The size of extracted supervoxels is impressive, superseding any manual skeletonization approaches. Further stages in the pipeline are anticipated to use tools such as Neuroglancer and CATMAID for proofreading, circuit analysis and visualization. Finally, I introduce a novel web-based platform for large-scale comparative connectomics – BrainCircuits.io. Initially, the platform indexes connectomics publications and links to publicly available image dataset for browsing. The platform is maintained and further developed by a newly formed company – UniDesign Solutions (https://unidesign.solutions) – which also provides additional connectomics services to the community.

### References

1 Titze B., Genoud C., and Friedrich R. W. (2018). *SBEMimage: Versatile Acquisition Control Software for Serial Block-Face Electron Microscopy.* Front Neural Circuits. 2018; 12: 54.

## 3.11 Commodity Connectomics and Applications to Bioinspired Robotics

*William Gray Roncal (Johns Hopkins Univ. – Baltimore, US)*



This talk gave an overview of the work related to large-scale connectomics done at the Applied Physics Lab (APL) Intelligent Systems Center at Johns Hopkins University, including the SABER pipeline, data storage, processing, and public outreach.

## 3.12 Dense connectomic reconstruction in layer 4 of the somatosensory cortex

*Moritz Helmstaedter (MPI for Brain Research – Frankfurt am Main, DE)*



The dense circuit structure of the mammalian cerebral cortex is still unknown. With developments in 3-dimensional (3D) electron microscopy, the imaging of sizeable volumes of neuropil has become possible, but dense reconstruction of connectomes from such image data is the limiting step. Here, we report the dense reconstruction of a volume of about 500,000 $\mu m^3$ from layer 4 of mouse barrel cortex, about 300 times larger than previous dense reconstructions from the mammalian cerebral cortex. Using a novel reconstruction technique, FocusEM, we were able to reconstruct a total of 0.9 meters of dendrites and about 1.8 meters of axons investing only about 4,000 human work hours, about 10-25 times more efficient than previous dense circuit reconstructions. We find that connectomic data alone allows the definition of inhibitory axon types that show established principles of synaptic specificity for subcellular postsynaptic compartments. We find that also a fraction of excitatory axons exhibit such subcellular target specificity. Only about 35% of inhibitory and 55% of excitatory synaptic subcellular innervation can be predicted from the geometrical availability of membrane surface, revoking coarser models of random wiring for synaptic connections in cortical layer 4. We furthermore find evidence for enhanced variability of synaptic input composition between neurons at the level of primary dendrites in cortical layer 4. Finally, we obtain evidence for Hebbian synaptic weight adaptation in at least 24% of connections; at least 35% of connections show no sign of such previous plasticity. Together, these results establish an approach to connectomic phenotyping of local dense neuronal circuitry in the mammalian cortex. [1]

### References

1 Motta, A., Berning, M., Boergens, K. M., Staffler, B., Beining, M., Loomba, S., Schramm, C., Hennig, P., Wissler, H, and Helmstaedter, M. (2018). *Dense connectomic reconstruction in layer 4 of the somatosensory cortex*. https://doi.org/10.1101/460618, https://www.biorxiv.org/content/early/2018/11/03/460618

## 3.13 Challenges for automated reconstruction

*Michal Januszewski (Google Switzerland – Zürich, CH)*



Significant advancements in automated reconstruction of neural tissue have been made in the last few years, yet much further progress is needed in order for the community to be able to tackle continuously increasing datasets. In my talk, I will try to highlight a number of issues that seem to be very important, but often do not receive sufficient attention. In particular, I will discuss the problems of acceptable error rates, practical proofreading time optimization and its relation to compute cost of automated processing, storage costs and interaction of compression methods with downstream processing, as well as alignment automation and data nonuniformity at large scales.

## 3.14 From electron microscopy images to a connectome

*Joergen Kornfeld (MPI für Neurobiologie – Martinsried, DE)*



Reconstruction and annotation of volume electron microscopy data sets of brain tissue is challenging but can reveal invaluable information about neuronal circuits. Significant progress has recently been made in automated neuron reconstruction as well as automated detection of synapses and the morphological classification of these reconstructions. Here, we present a complete computational analysis pipeline, starting with an aligned electron microscopy data set, and leading to a richly annotated connectivity matrix, including detailed semantic information about the location of dendritic spines and their presynaptic partners.

## 3.15 Whole organism segmentation without connectome

*Anna Kreshuk (EMBL – Heidelberg, DE)*



This talk described the efforts to do automatic segmentation analysis of cells in a large electron-microscopic dataset of a marine worm.

## 3.16 Sensorymotor circuitry in the fly peripheral nervous system

*Wei-Chung Allen Lee (Harvard Medical School – Boston, US)*



Wei showed data and results of his project to reconstruct the neuronal circuitry in the fly nerve cord from a large-scale EM volume and x-ray microCT data.

## 3.17 Mammalian Connectomes

*Jeff Lichtman (Harvard University – Cambridge, US)*



One looming challenge of connectomics is whether it will scale to volumes that comprise a mammalian brain. The mouse brain is about 500 cubic millimeters. If cut at a section thickness of 30 nm and imaged at 4 x 4 nm in-plane pixels, then this volume would require acquiring about 1 exabyte of data. This could be done in a number of ways but if the aim is to complete such a project in 5 years or less, all of them require parallelizing data acquisition. One way to do this is to section the volume first (such as on tape) and then distribute the imaging tasks by distributing the tape to multiple electron microscopes. To speed data acquisition further one can also parallelize the data acquisition from each microscope by use of multiple beam scanning devices (such as Zeiss' mSEM). For example, 24 multiple beam scanning electron microscopes each acquiring 42 TB per day could acquire an exabyte in 2 years 9 months. Storage of the data, stitching and alignment of the digital images, segmentation of cells and synapse and circuit reconstruction, and finally data sharing would require an extensive computer infrastructure and sophisticated AI technologies. Proof of concept computational pipelines for all these steps already exist in the Jain group at Google, Seung group at Princeton, and the Helmstaedter group at Frankfurt MPI among others. It thus seems feasible to do a project on this scale. A discussion ensued about how such a project might come about and what technical hurdles remain.

## 3.18 Ad hoc proofreading and analysis workflows using the Neuroglancer Python integration

*Jeremy Maitin-Shepard (Google Research – Mountain View, US)*



The Neuroglancer Python integration provides a way to use Neuroglancer as an interactive visualization library from Python, with the ability to overlay in-memory and on-the-fly generated volumes over existing volumes, define custom actions for keyboard and mouse events, and display annotations. Common use cases include manually annotating a set of objects/object fragments as one of several classes, displaying convolutional network inference results, visualizing the progress of a flood filling network-type algorithm, interactively viewing synaptic partner information, interactively splitting under-segmented objects using an agglomeration graph, and creating scripted videos.

## 3.19 Cross-Classification Clustering (3C): An Efficient Multi-Object Tracking Technique for 3-D Instance Segmentation in Connectomics

*Yaron Meirovitch (Harvard University – Cambridge)*



Pixel-accurate tracking of objects is a key element in many computer vision applications, often solved by iterated individual object tracking or instance segmentation followed by object matching. Here we introduce *cross classification clustering* (3C), a new technique that simultaneously tracks all objects in an image stack. The key idea in cross-classification is to efficiently turn a clustering problem into a classification problem by running a logarithmic number of independent classifications, letting the cross-labeling of these classifications uniquely classify each pixel to the object labels. We apply the 3C mechanism to achieve state-of-the-art accuracy in connectomics. Our reconstruction system introduces an order of magnitude scalability improvement over the best current methods for neuronal reconstruction, and can be seamlessly integrated within existing single-object tracking methods like flood-filling networks to improve their performance. This scalability is crucial for the real-world deployment of connectomics pipelines, as the best performing existing techniques require computing infrastructures that are beyond the reach of most labs.

## 3.20 Reconstructing subcellular microcircuits with circuit scale 3D electron microscopy

*Josh Morgan (Washington University, US)*



Many neurons have neurites that are both pre- and postsynaptic. These input/output neurites mean individual cells can contain multiple subcellular pathways for signal processing. Characterizing these subcellular pathways requires connectome analysis tools that take the precise spatial distribution of synapses into account. Applying this approach to the local interneurons of the mouse lateral geniculate nucleus reveals three types of output processes in the same neuron, each generating distinct synaptic motifs.

## 3.21 Dense Projectomes and Analysis of Connectomes

*R. Clay Reid (Allen Institute for Brain Science – Seattle, US)*



In the first part of this talk Clay described the idea to image dense projectomes, comprised of all myelinated axons running between different brain areas in whole brains, by confocal optical imaging using a fluorescent stain for myelin. In the second part he discussed the idea to use the similarity between real neuronal networks and machine learning models (e.g., convolutional networks) to analyze and understand the relationship between brain circuitry and function.

## 3.22 Whole-Brain Projectomes

*R. Clay Reid (Allen Institute for Brain Science – Seattle, US)*



In this talk Clay gave more detail about his proposal to analyze all myelinated projection axons in complete mammalian brains with light microscopy.

## 3.23 The value of connectomes in poorly explored species

*Kerrianne Ryan (Dalhousie University – Halifax, CA)*



The report of the connectome of the tadpole larva of the basal chordate *Ciona intestinalis* has enhanced our knowledge of its nervous system and its value as a biological model. The analysis of this connectome revealed features of synaptic organization, novel neurons and neuronal types, clarified sensory relay organization to the motor complex and enabled identification of homologous circuits with those of vertebrates. These analyses and explorations into the reported connectome have helped to promote this model species within the world of neuroscience, and have impacted the research in chordate neurobiology by providing testable hypotheses, and a comprehensive and detailed network of cells and connections. High throughput connectomics techniques can now be applied to test hypotheses and conditions in this developing nervous system alongside genetic tools.

## 3.24 Status report + Paintera

*Stephan Saalfeld (Howard Hughes Medical Institute – Ashburn, US)*



Stephan talked about his current projects, including 'N5', a hierarchical tensor storage API with backends to filesystem, cloud-storage, and HDF5 that allows parallel writing of chunked data [1]. He showed their synaptic cleft detection method on the complete Drosophila brain imaged with serial section transmission EM [2, 3], their new Drosophila brain atlas [4] bridging between light and EM, and 'Paintera' [5], a tool they are developing for manual image painting and proofreading on large data sets.

### References

1 https://github.com/saalfeldlab/n5
2 Heinrich, L., Funke, J., Pape, C., Nunez-Iglesias, J., and Saalfeld, S. (2018). *Synaptic Cleft Segmentation in Non-Isotropic Volume Electron Microscopy of the Complete Drosophila Brain.* In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, 317–25. https://doi.org/10.1007/978-3-030-00934-2_36.
3 Zheng, Z., Lauritzen, J. S., Perlman, E., Robinson, C. G., Nichols, M., Milkie, D., Torrens, O., et al. (2018). *A Complete Electron Microscopy Volume of the Brain of Adult Drosophila Melanogaster.* Cell 174, no. 3 (July 26, 2018): 730-743.e22.

4 Bogovic, J. A., Otsuna, H., Heinrich, L., Ito, M., Jeter, J., Meissner, G. W., Nern, A., et al. (2018). *An Unbiased Template of the Drosophila Brain and Ventral Nerve Cord.* BioRxiv, July 25, 2018, 376384. https://doi.org/10.1101/376384.
5 https://github.com/saalfeldlab/paintera

## 3.25 Ontogeny, phylogeny, and connectomics

*Aravinthan D.T. Samuel (Harvard University – Cambridge, US)*



Developmental biology and evolutionary biology are interwoven. Adaptive changes in an animal's anatomy and behavior, which confer fitness in an evolutionary sense, can occur as modulations of developmental programs. This is probably also true of neural circuits and brain connectivity, and could be studied through connectomics applied to animals across their developmental time courses and across their phylogenetic neighbors. Given the current pipelines for connectomics, it makes most sense to pursue this endeavor with small animals like nematodes and flies. Here, I argue for connectomics focused on small invertebrates to pursue such fundamental questions in biology.

## 3.26 Publishing and Simulation

*Louis Scheffer (Howard Hughes Medical Institute – Ashburn, US)*



In science, when you publish, the data should be available, and the analysis reproducible. This is challenging with large connectomics data sets. Currently we rely on the graciousness of the authors, but we should likely move to centralized analysis and storage, as genetics did long ago. Another use of completed connectomes is simulation. This seems simple, but in practice is much more than re-formatting. You must pick a subset of data, extrapolate to portions of neurons outside your volume, find synapse models, and figure out input vectors and observables. The raw models from EM are more complex than needed and make simulation slow without adding accuracy. We need fairly sophisticated model order reduction to make this practical.

## 3.27 The Dynamic Connectome

*Jochen Triesch (Goethe-Universität Frankfurt am Main, DE)*



Experiments from recent years suggest that the connectome is quite dynamic [1]. On the one hand, learning processes have been shown to induce systematic changes to the connectome. On the other hand, even under basal conditions there is substantial turnover of synaptic connections. Nevertheless, the connectome also has stable features such as the lognormal-like

distribution of postsynaptic density surface area, which is a good proxy for synaptic efficacy. I will discuss self-organizing recurrent neural network models (SORNs), that try to reproduce and explain the dynamic and stable features of the connectome. Interestingly, a recent model from this family can also capture experiments on sequence learning in rodent visual cortex and predicts specific learning-induced changes to the connectome. [2]

### References

1 Rumpel, S., & Triesch, J. (2016). *The dynamic connectome*. e-Neuroforum, 22(3), 48-53.
2 Klos, C., Miner, D., & Triesch, J. (2018). *Bridging structure and function: A model of sequence learning and prediction in primary visual cortex*. PLoS computational biology, 14(6), e1006187.

## 3.28 The network topology of neural sequences in retrosplenial cortex

*Adrian Wanner (Princeton University, US)*



Working memory, the ability to temporarily hold multiple pieces of information in mind for manipulation, is central to virtually all cognitive abilities. Preliminary data from the Tank lab at the Princeton Neuroscience Institute shows choice-specific activity sequences in posterior cortical neurons in a delayed match to sample task, which can be interpreted as working memory related neural activity. I aim to comprehensively dissect the neural circuit underlying these sequential activity patterns by combining *in-vivo* functional imaging in mice performing working memory tasks in a virtual-reality setting with subsequent large-scale electron microscopy (EM) based circuit reconstruction. Thereby I focus on retrosplenial cortex, which is of special interest because working memory related activity sequences are more linear in this part of cortex than in other cortical areas. The neurons that participate in these sequences are typically hundreds of microns apart from each other. Reconstructing the underlying microcircuit therefore requires the acquisition of cubic millimeter sized EM volumes. Staining such large volumes with high contrast homogeneously is still very challenging and unreliable. To facilitate the development of new staining protocols and to monitor the staining process in precious functionally characterized tissue blocks, I developed a novel X-ray assisted staining procedure [1], that uses temporally resolved high-resolution X-ray imaging to monitor the diffusion and staining of heavy metals into the tissue – enabling to control the staining process in much more detail. In collaboration with the Allen Institute of Brain Science in Seattle, we recently built a custom low cost high-throughput transmission electron microscopy (TEM) pipeline. Our pipeline features an automated reel-to-reel system for grid tape (Harvard University/Luxel Inc.) section feed [2] and a prototype of CRICKET, a TEM beam scanner developed by Voxa Inc. In combination with a novel 50 megapixel camera system from AMT Inc, that comes with a specialized integrated scintillator-lens system, we routinely reach acquisition rates above 250 MHz.

### References

1 Provisional patent application no. 62/760,329
2 Own, C., Murfitt, M., Own, L., & Cushing, J. (2017). *Developments in Reel-to-Reel Electron Microscopy Infrastructure*. Microscopy and Microanalysis, 23(S1), 32-33. doi:10.1017/S1431927617000848

## 3.29 Achieving the next order of magnitude in imaging speed with multibeam scanning electron microscopes

*Dirk Zeidler (Carl Zeiss – Oberkochen, DE) and Anna Lena Eberle (Carl Zeiss – Oberkochen DE)*



With the introduction of a 61-beam multibeam SEM in 2014, the limiting element in connectomics studies has shifted from data acquisition to data processing. As data processing capabilities have increased since, we have recently developed a 91-beam multibeam SEM that provides a 50% increase in throughput at similar resolution. These instruments are currently used for data acquisition at sample sizes of about 1 mm side length or below. Nevertheless, future large-scale studies with sample volumes of possibly several $cm^3$ will soon require higher throughput. We will report on latest results on the development of our multibeam technology. We will show results on recent resolution improvements at the latest 91-beam tool, and first results on our 331-beam demonstrator. We will also show roadmaps on reducing overhead and further improving resolution and scan speed.

# 4 Working groups

## 4.1 Alignment

*Working group participants*



Alignment and stitching of image stacks is one of the major steps in the pipeline of data processing for large-scale electron microscopy. This workgroup addressed current challenges and solutions. A typical alignment approach involves several stages such as: 1) Stitching montage tiles into 2D section. 2) Rough alignment of sections (Often SIFT to affine). 3) Fine non-linear warping. For small volumes (<1TB) ImageJ plugins and a standard workstation can adequately align. Very large volumes may require tile management software, cluster computing (or custom high powered workstations), and human supervision/correction. Alignment code is available from Github distributions from Stephan Saalfeld, Janelia, Sebastian Seung, Google and others. Notably, good alignment of large volumes still requires human expertise in tweaking parameters, human management moving data through the alignment pipeline, and human corrections of alignment errors. Both Forrest Coleman and Stephan Gerhard have alignment pipelines they are reasonably satisfied with. Forrest stressed the value of having managing tiles and transformations using the Render web service (from Janelia). Using this framework, a range of alignment solutions can be tested and integrated into the data processing pipeline.

## 4.2 Cell Types

*Working group participants*



Often, addressing biological questions using large electron-microscopic datasets involves the identification of different cell types. This workgroup discussed methods of identification, and the general question of what constitutes a 'cell type' in the first place. Cell types could, for example, be defined as the centers of clusters in a feature space spanning synaptic connectivity and morphology, a more flexible concept than the hopeless attempt of drawing sharp boundaries where none might exist. There was a general consensus that large-scale connectomics will allow us to focus more on individual neurons and how they interact with their synaptic partners, which should provide a better understanding of how similar neurons as individuals are.

## 4.3 Combining EM connectomics with non-EM techniques

*Working group participants*



This working group discussed what other techniques might be helpful to integrate with EM connectomes. The main division is between other techniques applied to the same animal, versus other techniques applied to other animals of the same type, but then cross-referenced to EM.

Techniques that can be applied in the same animal:

- Calcium (activity measurement) and other imaging
- X-ray of block
- Dye fill or other labelling such as nanobodies
- Electrophysiology (on same volume before prep)
- Labelled lines in same value

Techniques that can be applied across different animals:

- Corresponding genetic lines
- RNA expression data
- Papers/literature on cell type
- Links to other atlases
- Lineage and successors
- Gap junctions and other stuff not visible in EM

## 4.4 Error Metrics

*Working group participants*



Current machine learning methods can produce automatic segmentations of cells and their processes in 3D electron microscopic image datasets, but with non-zero error rates. Also, an error estimate ('loss function') is an important parameter used in training models in machine learning; often, training is done by adaptation of model parameters to minimize the error between the output of the model and the desired output ('ground truth'). Therefore, robust error metrics are needed to compare the performance of different automatic segmentation methods, and to train the segmentation algorithms.

Used metrics include: axon/dendrite run length, variation of information, Rand index, graph edit distance, and network integrity. There was some discussion about the differences between what was useful for a loss function vs. an error metric. Stefan Saalfeld argued that error metrics really aren't the core problem right now, and that having easy access to test data is more the limiting factor. He pointed out that techniques that work on specific data (i.e., CREMI challenge) are often failing on different data.

## 4.5 Full pipeline design

*Working group participants*



Currently the analysis of large electron microscopic data sets is done with a large variety of tools developed by many labs independently, each of which can deal with one or several parts of the whole analysis pipeline. This workshop discussed the possibility of combining efforts and coming up with a framework for tool integration, to reduce redundant programming efforts and to improve usability. It was felt that the pipeline should be developed and improved as an integrated product and software, and that it should be based on an open source framework, such as Fiji (ImageJ). However, even if such a framework existed, labs may still need a systems programmer to adapt it to their specific needs; so a 'full pipeline design' may have to consist of the combination of a software package and a hired programmer. Alternatively, it may be helpful if particularly useful features (like the 'flight mode' in Knossos) could be made available to the community and integrated in other tools.

## 4.6 Long-Term Storage

*Working group participants*



Current electron microscopic data sets can be very large (exceeding one Petabyte), which means that storage is expensive, especially if the data should be widely accessible and held for a long time.

There was no consensus about whether data should be made available beyond 5 years. Cost could be reduced by only storing compressed raw data, and it may in some cases be possible and cheaper to re-acquire the data if necessary. The usefulness of data for science could be quantified by a metric of # of scientists/megabytes. There should maybe be a public of notice when data is retired, or a defined data retiring plan. Ultimately, a timeline for storage and availability of the data produced in a scientific project could be specified in grant proposals, and the cost could be covered by the project's budget.

## 4.7 Proofreading Tools

*Working group participants*



Proofreading is one of the major parts of a typical data analysis pipeline of electron-microscopic image volume datasets. After segmentation of objects of interest with automatic methods, often human experts review and correct the result. This workgroup discussed current needs and currently available tools. These include:

- TrakEM2
- CATMAID
- Mojo/Dojo
- Knossos
- WebKnossos
- pyKnossos
- Eyewire
- NeuTu
- Raveler
- Ilastik
- Paintera
- VAST
- Neuroglancer

It was noted that as the quality of automatic segmentation results improves, finding errors may become inefficient because too much data has to be screened by experts to find errors. In that case targeted proofreading may help.

## 4.8 Raw Data Quality

*Working group participants*



The quality of image data acquired with electron microscopes can have profound effects on the more challenging data processing downstream, like image alignment and segmentation. For example, errors in automatic segmentation are often caused by defects in the image data. This workgroup addressed the need for image quality metrics to ensure optimal downstream processing.

Nobody could agree what the right quality was and intuitions collided. However, everybody agreed that we needed to know this and there was a consensus to collect a FIB (focused ion beam) dataset of mouse cortex at a resolution of 4 nm isotropic and 25 micrometer cube. Then this dataset would be degraded to different voxel sizes, anisotropies and signal to noise to estimate what would be the minimal conditions to collect a dataset.

There was some discussion about having reference volumes of varying section thickness, etc. Winfried Denk argued that it was going to be very hard to evaluate all those. There was some disagreement whether you can simulate different qualities by adding shot noise and simulating different aspect ratios. Some people aren't sure that would work. Clay Reid asked how much it would cost to acquire and how much to do ground truth segmentation of the dataset. People agreed that it would be money well spent.

Different researchers had different intuitions about what resolution is good enough. Sectioners feel comfortable with a section thickness of 25-40 nm. FIBers feel 10 nm is correct. However there is no data to draw upon yet, because ground truthing is boring and tedious and expensive.

## 4.9 Scalable analysis of connectomes, representation beyond graphs

*Working group participants*



This workgroup looked at the queries biologists would like to make of connectomes. Many of these involve information other than just the graph of connections. We listed a number of different queries and tasks users would like to perform. The numbers after each topic are number of people (of about 20) who would use this feature. Features marked with '##' need information beyond the graph:

- Link to other databases on the same species (All) ##
- Query by cell type
- List/table/query of all inputs and outputs
- Violations of Peter's rule (contact vs number/size of synapses) (5) ##
- Find pathways from A to B (all)
- Motifs, possibly local, beyond triplets (10)
- Shape based searches (all) ##
- Build realistic electrical models (4). Simulators include: Neuron (3) ##; Nest (1); Genesis (0); Brian (2); Dipty (2).
- Local fixed points (as determined by specific local motifs) (5)
- Contact area (all) ##
- Detection of bundles and assignment of neurons to bundles (all) ##
- Connectome differences, between connectomes and between subsets of same connectome (6).
- All the above, but: Interactive (6); Filterable (all).
- Staff, user community, etc. to get help on constructing efficient queries, etc. (6) ##
- Metadata curation (cell names, names of spines/necks, etc) (6) ##
- Ultrastructure information (subcellular segmentation). Run additional analysis on a selected subset (all) ##
- Confidence numbers (5)

- Correlation to functional imaging obtained beforehand (8) ##
- Biases in sampling (1)
- Overall metrics (density of neurons, synapses, branch points, etc) (all) ##
- Import into other tools (10) ##
- Complex queries on big systems paid for by the connectome project) (2) ##
- Running queries in the cloud (paid for by user): ##;
- – Should be possible, achieved by storing a copy in the cloud (all);
- – Should be easy (5)
- Natural language queries (4) ##
- Connection to proofreading (for double checking, looking for other cell characteristics, etc) (all) ##
- Log files and other methods to check operation of queries, find FAQ, etc. (6) ##

## 4.10 Synaptic Strength

*Working group participants*



Electron microscopy delivers static images of the structure of nervous systems, but not direct information of functional properties of the cells. Some parameters may however be inferred from the images. An important one is synaptic strength, which informs about the connection strength between neurons. Important image parameters include the size of the synaptic cleft and the number of neurotransmitter vesicles at the synapse.

In the fly Albert Cardona reported that in the larva there is a linear correlation between number of synapses and summed synapse area implying that all the synapses are the same size. Jeff Lichtman argued that this may change when an animal matures. Lou Scheffer reported that in the mushroom body there is variability of synapse size. He also reported that inactive/silent synapses do not disappear. A discussion followed on silent synapses in the mammalian brain, with no resolution, as we don't know if there are really silent synapses and if there are we don't know how they look at the EM level.

There was also an interesting discussion on electrical synapses. Kevin Briggman has data from the retina that shows if you prepare tissue with preservation of extracellular space, the appositions that are left and are not chemical synapses correlates with cell types that have electrical synapses. A discussion followed that it would be interesting to do the same in the cortex.

There was also a discussion of the roles of vesicle density and spine neck shape, without resolution.

## 4.11 Whole Brain Projects

*Working group participants*



As EM datasets now reach a scale of cubic millimeters of tissue and data sizes of Petabytes, it may soon become feasible to prepare and image larger brains in their entirety with electron-microscopic resolution. This workshop explored the feasibility of imaging whole mammalian brains, in particular a whole mouse brain.

# 5 Panel discussions

## 5.1 Final Discussion Transcript

*Seminar participants*



Winfried Denk thanked the organizers, most of who had left, for a meeting he enjoyed that much and which was that productive. Lou Scheffer said that he realized that people made serious thoughts about the whole mouse brain, which he thought to be 10-15 years away. Adrian Wanner stated that the atmosphere was really special, with new energy, more motivation, and people feeling more optimistic. Nir Shavit said that he thinks that we will have some graphs that we can analyze very soon. He thinks this community is fun, since people are working hand in hand towards the same end.

- Forrest: What makes things transfer? transmission of information and methods
- Will: Where does it make sense to compete and where does it make sense to collaborate
- Nir: Next Dagstuhl seminar?
- Winfried: One has always the desire to have the users/biologists, but it doesn't work, because of little common interest in each specific system. Once we reach the shores of graphs we will reach nirvana? But how to translate graphs into biological meaning? That may be the next difficult thing, and may make sense for a next meeting – theoretical neuroscientists and people who generate connectomes, even people who have no vested interest in neuroscience. Right now the methods discussion still dominates, but may be resolved in 2-3 years
- Nir: One could get a few of the top theorists in machine learning, it would be fascinating to expose them to things we know and they don't, to help them to incorporate biological insights. They view the connectome as a random graph, which is far away from reality, anchoring them to what a connectome really is
- Forrest: Connectome graph challenge for theorists? Solve practical problems on large graphs where the correct result is known?
- Joergen: Large-scale simulation community?
- Winfried: Question about challenges: Challenges provide a good endpoint, but in a phase when developing methods they may limit ideas, and may not work that well
- Winfried: We should focus on what they can do for us rather than what we can do for them, to develop our field. We need the input from the graph field, but modeling is not that useful for us
- Lou: Modeling can help us
- Winfried: Modelers will come whereever they can find data
- Nir: Challenge how to find motifs
- Jochen: Graph theorists don't think about dynamics of graphs
- Lou: Finding motifs in graphs with errors
- Winfried: Have we forgotten anything we need?
- Dirk: One direction: Small animal; use larger and larger animals; Other direction: Make portion of large brain larger. How to run large projects? How to talk to funding agencies? Talk to people who do that.
- Ask mathematicians for a new view on things, let in mathematicians, may solve a lot of problems by sitting there with a big cup of coffee and finding the small glitch in the method, a tiny glitch may have a large effect

- Winfried: We don't have enough to eat for a mathematician, we'd have to pay them
- Kerrianne: Some mathematicians already get very excited
- Winfried: Do we need people to run large projects in hierarchical settings (scary)? More important is a cottage industry of a number of small groups making progress on different problems, which may be in danger if there is a large project
- Joergen: Include the genomics part
- Winfried: Which approach advances the field more, the small projects developing different methods, or the big project? But 1 cubic millimeter is too small if one can't find the source of an axon of interest. Localized computation, but the cost function of behavior is global, so one has to be able to put it in a global context, which requires high resolution and global context
- Lou: That is the advantage of a small animal!
- Winfried: The fly work is crucial to make clear why connectomes are useful, which can be used as an argument
- Lou: Why is it that a child can learn what a tiger is in 3 examples, but machine learning takes a million?
- Winfried: The bane of modeling of the last 30 years is that you make assumptions and then something interesting happens and then you sit on this pile of rubble. Once you have the circuit diagram you can build theories on solid ground. Set the theorists free by constraining the theorists
- Julia: Should we have a large-scale segmentation benchmark? Maybe from two different techniques and two different model organisms, as a test case
- Lou: This needs a secret test set, so difficult to do
- Winfried: It would be ideal if my competitor did it, so they would be slowed down, but is it necessary for progress? We're not held back by the lack of this, usually
- Daniel: It could attract people from outside?
- Winfried: 7 years ago I would have agreed, but not necessary any more, there are enough talented people in the field
- Nir: MNist, imagenet have been extreme drivers of development; large datasets could be beneficial and would allow progress, also for the development of measures / metrics
- Jochen: Competition vs. benchmark: benchmarks are low-maintenance
- Michal: A volume size of 25 microns$^3$ is the lower bound
- Stefan: Computer vision conferences are full of talented people who are virtually unaware of our field; but yes, organizing such a challenge is major distraction
- Winfried: In the spirit of challenges, for the whole mouse brain: What preparation to choose, what imaging method to choose? Quantitative way – lowest error rate: All camps provide a reasonable volume with training data, and the decision to use one method or another is based on evidence rather than opinions. This requires thought to prevent to game this competition but is important. We should first analyze how error rates of different approaches compare.

## Participants

- Daniel R. Berger
Harvard University – Cambridge, US
- Davi Bock
Howard Hughes Medical Institute – Ashburn, US
- Kevin Briggman
Max-Planck-Gesellschaft – Bonn, DE
- Julia Buhmann
Universität Zürich, CH
- Albert Cardona
University of Cambridge, GB
- Forrest Collman
Allen Institute for Brain Science – Seattle, US
- Nuno Maçarico da Costa
Allen Institute for Brain Science – Seattle, US
- Winfried Denk
MPI für Neurobiologie – Martinsried, DE
- Eva Dyer
Georgia Institute of Technology – Atlanta, US
- Rainer W. Friedrich
FMI – Basel, CH
- Jan Funke
Howard Hughes Medical Institute – Ashburn, US
- Christel Genoud
FMI – Basel, CH
- Stephan Gerhard
FMI – Basel, CH
- William Gray Roncal
Johns Hopkins Univ. – Baltimore, US
- Moritz Helmstaedter
MPI for Brain Research – Frankfurt am Main, DE
- Michal Januszewski
Google Switzerland – Zürich, CH
- Joergen Kornfeld
MPI für Neurobiologie – Martinsried, DE
- Anna Kreshuk
EMBL – Heidelberg, DE
- Julia Kuhl
MPI für Neurobiologie – Martinsried, DE
- Wei-Chung Allen Lee
Harvard Medical School – Boston, US
- Jeff Lichtman
Harvard University – Cambridge, US
- Jeremy Maitin-Shepard
Google Research – Mountain View, US
- Yaron Meirovitch
Harvard University – Cambridge, US
- Josh Morgan
Washington University, US
- R. Clay Reid
Allen Institute for Brain Science – Seattle, US
- Kerrianne Ryan
Dalhousie University – Halifax, CA
- Stephan Saalfeld
Howard Hughes Medical Institute – Ashburn, US
- Aravinthan D.T. Samuel
Harvard University – Cambridge, US
- Louis Scheffer
Howard Hughes Medical Institute – Ashburn, US
- Nir Shavit
MIT – Cambridge, US
- Jochen Triesch
Goethe-Universität Frankfurt am Main, DE
- Xueying Wang
Harvard University – Cambridge, US
- Adrian Wanner
Princeton University, US
- Casimir Wierzynski
Intel – Santa Clara, US
- Dirk Zeidler
Carl Zeiss – Oberkochen, DE

Report from Dagstuhl Seminar 18482

# Network Visualization in the Humanities

**Edited by**

**Katy Börner[1], Oyvind Eide[2], Tamara Mchedlidze[3], Malte Rehbein[4], and Gerik Scheuermann[5]**

1 **Indiana University – Bloomington, US,** katy@indiana.edu
2 **Universität zu Köln, DE,** oeide@uni-koeln.de
3 **KIT – Karlsruher Institut für Technologie, DE,** mched@iti.uka.de
4 **Universität Passau, DE,** malte.rehbein@uni-passau.de
5 **Universität Leipzig, DE,** scheuermann@informatik.uni-leipzig.de

**Abstract**

This report documents the program and the outcomes of Dagstuhl Seminar 18482 "Network Visualization in the Humanities", which took place November 25–30, 2019. The seminar brought together 27 researchers from Network Visualization and Digital Humanities communities. During the seminar the participants shared knowledge on the existing methods of network visualization and on network visualization challenges present in the Humanities through the introductory talks, the abstracts of which are included in this report. Multiple innovative research challenges for Network Visualisation in the Humanities have been identified and according to those four working groups have been set up that discussed the topics in detail. The summary of the discussions of the working groups is given in this report.

**Seminar** November 25–30, 2018 – http://www.dagstuhl.de/18482
**2012 ACM Subject Classification** Human-centered computing → Visualization, Applied computing → Arts and humanities, Theory of computation → Design and analysis of algorithms
**Keywords and phrases** digital humanities, network visualization, graph drawing, human computer interaction, topic modelling, cyberinfrastructures, distant reading
**Digital Object Identifier** 10.4230/DagRep.8.11.139

## 1 Executive Summary

*Katy Börner (Indiana University – Bloomington, US)*
*Oyvind Eide (Universität zu Köln, DE)*
*Tamara Mchedlidze (KIT – Karlsruher Institut für Technologie, DE)*
*Malte Rehbein (Universität Passau, DE)*
*Gerik Scheuermann (Universität Leipzig, DE)*



### Seminar Goals

The application of computer-based methods by scholars of the Humanities has a tradition that goes back to the mid 20th century. Labelled "Digital Humanities" some 15 years ago, it has seen a significant growth since then [1]. An important part of Digital Humanities methodology is to establish data sets [2] based on cultural artefacts such as fiction texts, paintings, musical scores and recordings, and historical sources in all media. This is done in a number of different ways and includes some sort of extraction of data from sources


Network Visualization in the Humanities, *Dagstuhl Reports*, Vol. 8, Issue 11, pp. 139–153
Editors: Katy Börner, Oyvind Eide, Tamara Mchedlidze, Malte Rehbein, and Gerik Scheuermann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

structured in different, less explicit ways than what is needed for operationalisation and computer assisted analysis and visualization. When this process works well, it supports scholars' endeavours to answer existing research questions and to generate new insights and novel research questions. A significant part of the data collected can be modeled as networks.

Existing network analysis and visualization techniques have already proven themselves immensely useful in analyzing Digital Humanities data and providing new discoveries [3]. The central goal of research on network visualization for digital humanities scholars is to develop visualization techniques and algorithms that empower scholars to use those effectively as part of their research process and for communicating study results to readers. While network science approaches are widely used in other research areas, the power of a network mindset and approach has not yet been fully exploited within the Humanities.

The seminar aimed to enhance the development of network visualization algorithms and tools centered around humanities research. In particular, its goals were as follows:

- Interdisciplinary Exchange: to discuss existing network visualization methods and algorithms in perspective of their potential application within the Humanities;
- Terminology Gap: Bridging the gap in terminology between Digital Humanities on one side and computer scientists in Network Visualization and Graph Drawing on the other side;
- Data: to discuss Humanities' data sources and their nature, research questions, use cases, and specific application profiles in perspective of their potential support by network visualisation.
- Reserch Agenda: Formulation of research agenda on "Network Visualization in the Digital Humanities". Creation of interdisciplinary teams of researchers that address specific scientific challenges of the agenda;

## Seminar Program

The seminar brought together 27 researchers from Network Visualization and Digital Humanities communities. The initial two days of the weeklong event were devoted to bring together the different communities and to develop a mutual understanding. Researchers informed each other about their scholarly background through short, five-minute talks. In addition, there were eight long, 45 minutes, presentations in which digital humanities scholars discussed network and network visualization challenges and opportunities in their field of expertise. This was complemented by surveys on network visualization and successful examples of cooperation between visualization and digital humanities researchers.

During both days the participants were asked to post questions and issues they would like to discuss in the remaining three days of the seminar. After a voting, four research areas most interesting the participant were identified. All four met the guiding principles in that they describe both: highly relevant applications within the Humanities as well as innovative research challenges for Network Visualisation. They are as follows:

- Complex networks, in particular multivariate, multilayered, and multilevel networks;
- Linked networks;
- Temporal networks;
- Uncertainty, incompleteness, and ambiguity of data.

Four groups were formed to work on those four topics over the remaining three days. There were several opportunities for joint discussions and progress reports across the groups. Summaries of the group discussions can be found in Section 4.

## Future Plans

During the seminar the participants decided to proceed with a publication of a manifesto, outlining a research agenda for "network visualisation in the Humanities". It was also planned to publish an edited volume on specific aspects of the overarching topic, possibly along the four major research areas identified by the seminar. The volume will be submitted as a special issue to "Historical Network Research", an Open Access Journal.

## Evaluation

The feedback provided by the participants in form of a survey collected by Schloss Dagstuhl was highly positive and in most aspects above the average collected over the last 60 seminars. The participants agreed that the seminar inspired new ideas, collaborations, joint publications and brought insight from neighboring fields. There was a number of positive comments by the participants on the structure and organization of the seminar as well as several useful suggestions for the future seminars.

## Acknowledgment

As an organizing committee of the seminar we would like to thank the scientific and administration staff of Schloss Dagstuhl for the excellent support they provided, both in the preparation phase and during the seminar. On behalf of all participants, we would also like to thank Dagstuhl for the high quality facilities provided, for excellent rooms for work and socializing, for the tasty meals, and of course also for the excellent wine cellar. The organizers of the seminar would also like to thank Ray Siemens and Dan Edelstein for their contributions to the initial Dagstuhl proposal. Finally, we thank Christina Gillmann for taking the responsibility for this report.

### References

1 Thaller, Manfred (2017), *Geschichte der Digital Humanities.* In: Fotis Jannidis, Hubertus Kohle und Malte Rehbein (Hg.): *Digital Humanities. Eine Einführung.* Stuttgart, J.B. Metzler Verlag, S. 3–12.

2 Schöch, Christof (2013) *Big? Smart? Clean? Messy? Data in the Humanities.* In: Journal of Digital Humanities 2 (3).

3 Rehbein, Malte (forthcoming) *Historical Network Research, Digital History, and Digital Humanities.* In: Marten Düring, Florian Kerschbaumer, Linda von Keyserlingk und Martin Stark (Hg.): *The Power of Networks. Prospects of Historical Network Research*, Routledge.

## 2 Contents

# 3 Overview of Talks

## 3.1 Almost a Theory

*Ulrik Brandes (ETH Zürich, CH)*



Data, at least in the statistical sense, are the values of variables, which we conceive of as mappings of entities from some domain to values from some range. The distinct characteristic of network variables is that their domain consists of overlapping dyads. While the representational theory of measurement is concerned with the preservation of empirical structures on the domain in numerical structures on the range, a particular challenge in the Humanities is the conceptualization of the domain. Many relevant aspects of the phenomena or artifacts under scrutiny are difficult to represent in variables because abstraction is necessary to make them commensurable. With increasing levels of abstraction, comparability is widened at the expense of potentially relevant characteristics. This is reflected, for instance, in the notions of close and distant reading, and suggests interesting problems for network visualization research.

## 3.2 Actionable Data Visualizations

*Katy Börner (Indiana University – Bloomington, US)*


**Main reference** Katy Börner: "Atlas of Knowledge: Anyone Can Map", Cambridge, MA: The MIT Press, 2015.
**Main reference** Katy Börner, David E. Polley: "Visual Insights: A Practical Guide to Making Sense of Data", Cambridge, MA: The MIT Press, 2014.
**Main reference** Katy Börner: "Atlas of Science: Visualizing What We Know", Cambridge, MA: The MIT Press, 2010.

In the information age, the ability to read and make data visualizations is as important as the ability to read and write. This talk explains and exemplifies the power of data visualizations not only to help locate us in physical space but also to help us understand the extent and structure of our collective knowledge, to identify bursts of activity, pathways of ideas, or emerging areas of research. It introduces a theoretical visualization framework meant to empower anyone to systematically render data into insights together with tools that support temporal, geospatial, topical, and network analyses and visualizations. Materials from the Information Visualization MOOC (http://ivmooc.cns.iu.edu) and science maps from the Places & Spaces: Mapping Science exhibit (http://scimaps.org) will be used to illustrate key concepts and to inspire participants to visualize their very own data.

## 3.3 Modeling Data to Develop Intuitions

*Nicole Coleman (Stanford University, US)*



**Main reference** Johanna Drucker: "Humanities Approaches to Graphical Display", Digital Humanities Quarterly, Vol. 5(1), 2011.
**URL** http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html
**Main reference** Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, Nicole Coleman: "Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project", The American Historical Review, Volume 122, Issue 2, pp. 400–424, 2017.
**URL** https://doi.org/10.1093/ahr/122.2.400
**Main reference** Stefano Franchi: "The Past, Present, and Future Encounters between Computation and the Humanities", pp. 349–364, Springer, 2013.
**URL** http://dx.doi.org/10.1007/978-3-642-31674-6_26

A decade of problem-driven experiments with data visualization for humanities research at Humanities + Design research lab at the Center for Spatial and Textual analysis at Stanford University exposed challenges and opportunities at the intersection of humanistic inquiry and data visualization. The challenges result from the fact that humanistic inquiry tends to be a very internal thought process that does not explicitly reference external models. To externalize the research process it was necessary to learn to think procedurally, to think visually, and to capture the complexity of multiple perspectives. This talk elaborates on a number of preliminary visualization experiments, which eventually led to three opensource applications: Palladio, Breve, and Data Pen. There were mistakes and successes along the way to understanding how to produce visualization tools that reflect the needs of humanistic research. The result is a set of tools that reveal the incompleteness of data, that move seamlessly from abstraction back to rich contextual sources, that allow reflection through interaction, and that allow visual data modeling, including direct editing and enrichment of a data set. The principles underlying the design of the interaction environments we built are analogous to those within the artificial intelligence community that support the augmentation of human ability rather than autonomous systems. Computation manages the complex calculation and querying of multi-dimensional data; the visual interface renders the data concisely and intuitively; and the interaction makes the visualization an exploratory tool. This combination, with an even emphasis on missing and extant data, provides an instrument for modeling data to develop intuitions. It is a fundamental functional design of data-driven environments that can be applied to far more sophisticated underlying computational and visual techniques to augment humanistic inquiry. Finally, to legitimize this new form of scholarship means to make it publishable. The question remains, how can we share the interactive experience of these works in a way that will persist and contribute to future knowledge production?

## 3.4 Visualization of networks: A cartographer's view

*Sara Irina Fabrikant (Universität Zürich, CH)*



Maps have been used for at least 5000 years by humans to communicate about tangible networks (i.e., transportation, water, electricity, etc.) and intangible linear features of the environment (i.e., flows of goods, people, air masses, etc.). Networks can represent geographic, linear features in the environment, or metaphorical, relational information spatialized from

databases that are not necessarily spatial or geographic (i.e., relationships extracted from text archives, biological databases, financial records, etc.). Spatialized displays can be designed and explored as if they represented geographic information, considering the user, the use context, and the design characteristics of the spatialized views. Empirical studies on spatialized views depicting points, networks, and regions suggest that cartographically informed design guidelines allow information seekers to more effectively and efficiently explore relational information, and gain knowledge from large spatialized text databases.

## 3.5 Visualization & Digital Humanities

*Stefan Jänicke (Universität Leipzig, DE)*



Visualization as a research instrument for Digital Humanities scholarship has gained more and more importance in recent years, yielding mutual benefits for the involved research domains. On the one hand, visualizations aim at providing intuitive access to vast amounts of digitized data, on the other hand, the nature of digital humanities data, which is usually incomplete, inhomogeneous and uncertain, provokes new challenges for visualization research. The best practice to develop a visualization in a digital humanities project is a problem-driven user-centered design approach as it forces scholars from both fields to intensely engage with each others research interests, tasks and problems. Moreover, the strong interdisciplinary exchange increases the likelihood that the resultant visualization will serve the intended purpose, and that it prepares the ground for future research.

## 3.6 Graphs in Computational Literary Studies

*Fotis Jannidis (Universität Würzburg, DE)*



Graphs are used for very different purposes in Computational Literary Studies and the talk discusses three different examples. 1) In a complex digital edition of Goethe's Faust the statements of researchers about the chronological relations (edges) between manuscripts (nodes) allow us to find contradicting statements. 2) In Stylometry the stylistic similarity (edges) between texts (nodes) can be used to determine groups beyond authorship based on period or other factors. 3) In the analysis of plot events (nodes) are related chronologically (edges). In the last case more complex graphs, which would allow to render narratological concepts like the difference between discours/histoire, are discussed. Modeling novels like that would allow us to find similar texts based on the similarity of the graphs representing their plot structure.

## 3.7 Putting Networks on the Map

*Stephen Kobourov (University of Arizona – Tucson, US)*



**Main reference** Randy Burd, Kimberly Andrews Espy, Md. Iqbal Hossain, Stephen G. Kobourov, Nirav Merchant, Helen C. Purchase: "GRAM: global research activity map", in Proc. of the 2018 International Conference on Advanced Visual Interfaces, AVI 2018, Castiglione della Pescaia, Italy, May 29 – June 01, 2018, pp. 31:1–31:9, ACM, 2018.
**URL** http://dx.doi.org/10.1145/3206505.3206531

Relational data sets are often visualized with graphs: objects become the graph vertices and relations become the graph edges. Graph drawing algorithms aim to present such data in an effective and aesthetically appealing way. We describe map representations, which provide a way to visualize relational data with the help of conceptual maps as a data representation metaphor. While graphs often require considerable effort to comprehend, a map representation is more intuitive, as most people are familiar with maps and standard map interactions via zooming and panning. Map-based visualization allows us to explicitly group related vertices into "countries" and to present additional information with contour and heatmap overlays. We discuss the graph-to-map (GMap) algorithmic framework, including applications, such as Maps of Computer Science (MoCS) and the Global Research Activity Map (GRAM), as well as experimental results on the effectiveness of the approach.

## 3.8 Networks in the Humanities: Challenges & Opportunities

*Scott Weingart (Carnegie Mellon University – Pittsburgh, US)*



Networks have a long history in the Humanities, going back to the earliest sociometry research in the 1930s, with contributions flowing both directions. By 2010, networks became a pillar of digital humanities research. The renewed interest brought to light several broad challenges needing to be addressed over the next decade, including toolkits that account for uncertain data, that support the entire research workflow, and that prioritize readability. The Humanities ought to also contribute new theoretical and historical understandings of network visualizations and analyses.

# 4 Working groups

## 4.1 Network Taxonomies

*Oyvind Eide (Universität zu Köln, DE), Francis Harvey (Leibniz Institut für Länderkunde – Leipzig, DE), Andreas Kerren (Linnaeus University – Växjö, SE), Tamara Mchedlidze (KIT – Karlsruher Institut für Technologie, DE), and Florian Windhager (Donau-Universität Krems, AT)*



During the meeting, our group met and discussed network taxonomies in the Humanities, a broad and intriguing topic because of its relevance for situating network research and development in the field. Correspondingly, the scope of the discussion ranged over the breadth and depth of network-based analysis in the Digital Humanities as well as in the Humanities at large. We recognized that taxonomies of networks in the Humanities are challenging to develop because of their heterogeneity. From this starting point we made progress by considering the disciplinary roots of networks and identified differences between types of networks. These points were then taken up when we focused on different networks used in the Digital Humanities, the data they were based on, and visualization techniques. We recognized how important terminological standardization is, and suggested it can be accomplished to some degree through taxonomies. Specific topics the group took up include: scale, interaction, hierarchy of networks, types of networks, including multivariate networks, multilayer networks, multimodal networks, and types of visualisations. Uncertainty in models and data as well as the challenge of accounting for the open nature of digital humanities research entered into the discussion as key points to be considered when developing taxonomies. The support of multiple analytical perspectives on data representations and interactive visual representations is necessary in order to promote and support the potential of networks for digital humanities research. Addressing the heterogeneity of networks and research theories, methodologies and context is necessary to account for the richness and unique composition in humanities research. This reminded us that our taxonomies can be helpful for a number of reasons and can fulfill different needs. The potential for standardization of terms is considerable, but taxonomies should not end with attempts at canonical specification. This could constrain scientific concept development. Taxonomies need to be dynamic and develop over time in order to keep them helpful and useful in the development of research.

## 4.2 Linked Networks Perspectives for Humanities (beyond Nodes, Links and Clusters)

*Gregor Betz (KIT – Karlsruher Institut für Technologie, DE), Stephen G. Kobourov (University of Arizona – Tucson, US), Martin Nöllenburg (TU Wien, AT), Gerik Scheuermann (Universität Leipzig, DE), Timothy Tangherlini (University of California at Los Angeles, US), and Christopher Warren (Carnegie Mellon University – Pittsburgh, US)*



The members of the linked networks perspectives for Humanities working group were Gregor Betz, Stephen Kobourov, Martin Nöllenburg, Gerik Scheuermann, Timothy Tangherlini, and Christopher Warren. We discussed three main questions: different networks based on the same data due to varying aspects of interest, good visualizations for bipartite/k-partite networks and how to draw linked networks. We see a clear need for research on these questions based on the needs in the Humanities and lack of approaches on the computer science side.

With respect to different networks based on the same data, we discussed several cases where there are good reasons to create different networks based on the same data, even for the same question. For example, one might consider a play with several scenes. For each scene, one can construct a network of the interactions of the characters. Then, the humanities scholar is left with the task of comparing these networks with respect to differences or consensus. Also, a very important task is the critical review of the network construction that questions all definitions of nodes or edges by looking at the original text. Especially the last task is not supported by current network visualization tools.

Regarding bipartite and k-partite graphs, we detected a high need of such graphs in the Humanities. E.g. any analysis of a novel that is based on the relationship between e.g. characters and scenes, character groups, characters and locations, etc. leads to such graphs. Surprisingly, there is only very little work in the literature on drawing bipartite graphs besides the common notion of drawing each set on one side of the screen. For k-partite graphs, there is even less work. This defines a clear need on the Network Visualization and Graph Drawing side.

As third part, we discussed the drawing of linked networks. Humanities scholars create usually several networks to describe relations of interest in their study. Quite often, they also create or note links between nodes (sometimes edges) in these networks. An example is korean pop music. One might look at the performers and their grouping into bands. One can also study the relations to a (fairly small) group of music production companies. In addition, one may study the role of auxiliary contributors such as choreographers, songwriters or promoters. This creates interlinked networks. The drawing and visual analysis of such linked networks is not well researched so far and presents a lot of challenges to the Graph Drawing and Network Visualization communities.

Overall, we found in all three parts substaniell potential for new solutions from computer science and high research interest from the Humanities. We expect fruitful cooperation among the participants of the seminar and beyond on this topic and encourage everyone interested to start research. Of course, the group is willing to share further thoughts on request.

## 4.3 Visualizing Networks and Temporality

*Melanie Conroy (University of Memphis, US), Kimmo Elo (University of Turku, FI), Gerhard Heyer (Universität Leipzig, DE), Fotis Jannidis (Universität Würzburg, DE), Malte Rehbein (Universität Passau, DE), Antonis Symvonis (National TU – Athens, GR), and Scott Weingart (Carnegie Mellon University – Pittsburgh, US)*



The members of the temporality working group were Melanie Conroy, Kimmo Elo, Gerhard Heyer, Fotis Jannadis, Malte Rehbein, Antonic Symvonis, and Scott Weingart. We discussed both ways to graph temporal networks and how to visualize data in the Humanities that include networks. While we discussed problems which could be addressed using temporal graphs, we found that temporal graphs could often not be constructed from the datasets with which we were familiar. Our discussion focused on how to incorporate non-linear time sequences and ways of perceiving time that differ from chronological time into network diagrams and other visualisations. After attempting to construct network graphs for various use cases, we discovered that many problems in the Humanities do not permit the construction of a temporal network graph due to multiple perspectives on the network and variable or uncertain time sequences. While we discussed ways to reduce the number of perspectives and series, however, we rejected the idea of reducing the complexity of data models. We decided that starting with the visualizations that we needed for a number of case studies would be more valuable than attempting to reduce the complexity of humanities research questions to make them graphable as a single network. For this reason, we decided to work backwards from the types of visualisations needed for individual projects to the data model that would be necessary to produce such a visualisation.

Problems that appeared repeatedly in our discussions of the temporality of networks in the Humanities included a mismatch between methodology (data models and metrics) and available technology, different data models and collection practices in various humanities fields, and project-specific data models. We also discussed incomplete or uncertain data and shifting or incommensurable perspectives related to time. We decided that no one visualisation or set of visualisations would be adequate to deal with all of these issues.

We discussed four main use cases for networks in the Humanities:

1. Story vs. Discourse – Literary character networks, in which the nodes are literary characters and the edges are co-occurrences in a series of scenes.
2. Word Co-Occurrence – Evolution of word use over time (word careers), in which the nodes are words and co-occurence in a text is represented by the edges.
3. Republic of Letters – Correspondence networks, in which the nodes are correspondents and the edges are letters.
4. Reports of Secret Police – Network model of the evolving knowledge of investigators into the relations of conspirators, in which each agent has a different view on the network of possible conspirators and nodes in the network appear and disappear as the police discover more about the network.

One idea that recurred frequently in our discussion was “snapshots” of a network which could be arranged into series by linking them to produce sequences instead of graphing a single temporal network. We designed and refined visualizations which could be used in each of these cases. Solutions included a stream graph of centrality and centralization, dyad visualization, temporal / witness matrix, collation networks, and a discourse/story/perspective model of

networks. Our solution to the problems presented by the variety and complexity of humanities data models was to combine network visualizations with representations of how the data was modeled–for example, placing network diagrams in a matrix that shows both the state of the network over time and how the network appears according to various perspectives which are made explicit in witness reports. By using a matrix, for example, we can show the state of a network across time according to various perspectives, such as witnesses to a series of events, or changes in the network.

For all four of our use cases, the combination of multiple visualisations was necessary to convey the most significant information about how the network was structured and how it developed over time; these visualisations could include a timeline or scatterplot to show the place in the temporal sequence of the network currently being visualized.

## 4.4 Uncertainty Visualization in Digital Humanities (DH) Network Data

*Katy Börner (Indiana University – Bloomington, US), Nicole Coleman (Stanford University, US), Marten Düring (University of Luxembourg, LU), Tim Dwyer (Monash University – Caulfield, AU), Sara Irina Fabrikant (Universität Zürich, CH), Christina Gillmann (TU Kaiserslautern, DE), and Stefan Jänicke (Universität Leipzig, DE)*



Uncertainty visualisation has been studied extensively in computer science, cartography, geography, information science, and related disciplines [1, 2, 3, 4]. Today, many individual solutions exist (e.g. [1, 2, 3]) within Digital Humanities (DH) projects but these have not yet been translated into generalizable solutions specific to either DH network data or the graph drawing community in general. Group members have diverse disciplinary backgrounds in computer science/visual analytics, information science, library studies, history, and GIScience/geovisual analytics. Such diversity within a group can be considered an advantage as it promises synergies between different conceptual frameworks. At the same time, it requires a shared terminology to avoid misunderstandings. Our team therefore drafted definitions of related key concepts such as error, uncertainty, graphs/network, sample bias, data quality, (in)homogeneity, probabilistic networks, data operationalization, and (in)compatibility with the goal to have them validated by the rest of the group. The representation of uncertainty in network visualisations in general and with regard to works in DH in particular, remains an open practical and research challenge. While specific problems such as the positioning of nodes due to uncertain node [5] or edge attributes [4] has received some attention in the past, to date there is no comprehensive overview of problems and recommended solutions. We expect that visual representations of uncertainty which were developed in other disciplines like information science, visual analytics, cartography, or bioinformatics can be adapted. This, however, raises the question to which extent data and research interests in DH differ from those in other domains. Is DH data special? The organizers of the Vis4DH workshop series [5] point to three distinctions: – differences in rhetorics of proof and discovery (and so differences in data culture and use) in the Humanities, – the difficulty of performing task analysis and evaluation for many humanities questions, that may have no ground truth, and finally, – in text visualization specifically, the difference between the needs of digital humanists (who perform close readings and critical engagements with texts) as opposed to

more standard text visualization scenarios (e.g. text analytics on datasets media analysis).

Here we seek to expand these observations further:

1. The data model itself is rhetorical in humanities research. Data modeling is part of an argument to be debated within the field. Data models therefore tend to be project-specific. So far, there are no generic data models which find general acceptance and are used for research across the DH.
2. Humanities data can be characterised by missing data, inhomogeneous representations of the available information and hard-to-resolve ambiguities. These problems appear as known unknowns and unknown unknowns.
3. Data visualisation is considered a (complementary) part of a research workflow alongside more traditional practices in DH.
4. Because historical datasets are inherently constructed, research data may include information from multiple sources, including attributes and values generated by the scholars. Data is seen to only partially represent scholarly knowledge. Interpretation requires enrichment with external information which defies representation in data.
5. Data analyses focus on data visualization as heuristic rather than as proof.
6. Data is often extracted or enriched manually which also explains the typically limited size of research datasets. Scholars seek to preserve the ability to manually edit, annotate and manage versions of their datasets.

This emphasis on often but not exclusively qualitative research practices, interpretation and personalised data models which also require representations of uncertainty [12] stands in contrast to the more empirical, probabilistic models of uncertainty which are common in other disciplines. A preliminary survey of interaction models developed in visual analytics [5, 6, 7, 8, 9, 10, 11] reveals that points 1 and 2 are not sufficiently captured in the existing approaches.

In order to properly define uncertainty visualisation challenges, we identified a preliminary taxonomy of different types of uncertainty in network visualisation. Existing taxonomies developed in other domains (cite) were reviewed and considered to be promising starting points. We identified four types of uncertainty (and their combinations) which we encounter frequently in network data:

- Time: When was it?
- Location: Where was it?
- Topic: What was it?
- Relation: What is the relation?

Our forthcoming survey paper will describe the state-of-the-art in depicting uncertainty in network visualisations, identify key challenges for Digital Humanities applications, and point towards best practices in cognate disciplines to resolve them.

## References


1 G. Caviglia and N. Coleman, *Idiographic Network Visualizations* Leonardo, vol. 49, no. 5, pp. 447-447, Mar. 2016.

2 S. Jänicke and D. J. Wrisley, *Visualizing Uncertainty: How to Use the Fuzzy Data of 550 Medieval Texts?*, in Conference Abstracts of the Digital Humanities 2013, 2013.

3 P. Girard et al., *RICardo Project?: Exploring XIX Century International Trade*, HAL, Post-Print hal-01835245, Jul. 2016.

4 J. Schwank, S. Schöffel, J. Stärz, and A. Ebert, *Visualizing Uncertainty of Edge Attributes in Node-Link Diagrams*, in 2016 20th International Conference Information Visualisation (IV), 2016, pp. 45-50.

5 D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, *The Role of Uncertainty, Awareness, and Trust in Visual Analytics*, IEEE Trans. Vis. Comput. Graph., vol. 22, no. 1, pp. 240-249, Jan. 2016.

6 T. Zuk and S. Carpendale, *Theoretical analysis of uncertainty visualizations*, presented at the Electronic Imaging 2006, San Jose, CA, 2006,

7 Torre Zuk, *Visualizing Uncertainty* Unpublished, 2008.

8 N. Andrienko et al., *Viewing Visual Analytics as Model Building* Comput. Graph. Forum, vol. 37, no. 6, pp. 275-299, 2018.

9 *(PDF) The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis* Online available: https://www.researchgate.net/publication/215439203_The_sensemaking_process_and_leverage_points_for_analyst_technology_as_identified_through_cognitive_task_analysis. Accessed: 29-Jan-2019.

10 *(PDF) Beyond Tools: Visual Support for the Entire Process of GIScience*, Online available: https://www.researchgate.net/publication/200043275_Beyond_Tools_Visual_Support_for_the_Entire_Process_of_GIScience. Accessed: 29-Jan-2019.

11 M. El-Assady et al., *Visual Text Analytics in Context of Digital Humanities*, Published at the 1st IEEE VIS Workshop on Visualization for the Digital Humanities as part of the IEEE VIS 2016, 2016.

12 *nodegoat | Blog | Formulating Ambiguity in a Database*, nodegoat, Online available: https://nodegoat.net/blog.s/21/formulating-ambiguity-in-a-database. Accessed: 19-Feb-2019.

## Participants

- Gregor Betz
KIT – Karlsruher Institut für Technologie, DE
- Katy Börner
Indiana University – Bloomington, US
- Ulrik Brandes
ETH Zürich, CH
- Nicole Coleman
Stanford University, US
- Melanie Conroy
University of Memphis, US
- Marten Düring
University of Luxembourg, LU
- Tim Dwyer
Monash University – Caulfield, AU
- Oyvind Eide
Universität zu Köln, DE
- Kimmo Elo
University of Turku, FI
- Sara Irina Fabrikant
Universität Zürich, CH
- Christina Gillmann
TU Kaiserslautern, DE
- Hans Hagen
TU Kaiserslautern, DE
- Francis Harvey
Leibniz Institut für Länderkunde – Leipzig, DE
- Gerhard Heyer
Universität Leipzig, DE
- Stefan Jänicke
Universität Leipzig, DE
- Fotis Jannidis
Universität Würzburg, DE
- Andreas Kerren
Linnaeus University – Växjö, SE
- Stephen G. Kobourov
University of Arizona – Tucson, US
- Tamara Mchedlidze
KIT – Karlsruher Institut für Technologie, DE
- Martin Nöllenburg
TU Wien, AT
- Malte Rehbein
Universität Passau, DE
- Gerik Scheuermann
Universität Leipzig, DE
- Antonis Symvonis
National TU – Athens, GR
- Timothy Tangherlini
University of California at Los Angeles, US
- Christopher Warren
Carnegie Mellon University – Pittsburgh, US
- Scott Weingart
Carnegie Mellon University – Pittsburgh, US
- Florian Windhager
Donau-Universität Krems, AT