



DAGSTUHL REPORTS

Volume 4, Issue 7, July 2014

Feature Interactions: The Next Generation (Dagstuhl Seminar 14281) <i>Sven Apel, Joanne M. Atlee, Luciano Baresi, and Pamela Zave</i>	1
Crowdsourcing and the Semantic Web (Dagstuhl Seminar 14282) <i>Abraham Bernstein, Jan Marco Leimeister, Natasha Noy, Cristina Sarasua, and Elena Simperl</i>	25
Information-Centric Networking 3 (Dagstuhl Seminar 14291) <i>Dirk Kutscher, Taekyoung Kwon, and Ignacio Solis</i>	52
Network Attack Detection and Defense: Securing Industrial Control Systems for Critical Infrastructures (Dagstuhl Seminar 14292) <i>Marc Dacier, Frank Kargl, Hartmut König, and Alfonso Valdes</i>	62
Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities (Dagstuhl Seminar 14301) <i>Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler</i>	80
Digital Palaeography: New Machines and Old Texts (Dagstuhl Seminar 14302) <i>Tal Hassner, Robert Sablatnig, Dominique Stutzmann, and Ségolène Tarte</i>	112

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

December, 2014

Bibliographic information published by the Deutsche Nationalbibliothek

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

License

This work is licensed under a Creative Commons Attribution 3.0 Unported license: CC-BY.



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

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

The copyright is retained by the corresponding authors.

Aims and Scope

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

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

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

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

Editorial Board

- Susanne Albers
- Bernd Becker
- Karsten Berns
- Stephan Diehl
- Hannes Hartenstein
- Stephan Merz
- Bernhard Mitschang
- Bernhard Nebel
- Han La Poutré
- Bernt Schiele
- Nicole Schweikardt
- Raimund Seidel (*Editor-in-Chief*)
- Michael Waidner
- Reinhard Wilhelm

Editorial Office

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

Contact

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

Digital Object Identifier: 10.4230/DagRep.4.7.i

Feature Interactions: The Next Generation

Edited by

Sven Apel¹, Joanne M. Atlee², Luciano Baresi³, and Pamela Zave⁴

1 University of Passau, DE, apel@uni-passau.de

2 University of Waterloo, CA, jmatlee@uwaterloo.ca

3 Politecnico di Milano, IT, luciano.baresi@polimi.it

4 AT&T Labs Research, US, pamela@research.att.com

Abstract

The feature-interaction problem is a major threat to modularity and impairs compositional development and reasoning. A feature interaction occurs when the behavior of one feature is affected by the presence of another feature; often it cannot be deduced easily from the behaviors of the individual features involved. The feature-interaction problem became a crisis in the telecommunications industry in the late 1980s, and researchers responded with formalisms that enable automatic detection of feature interactions, architectures that avoid classes of interactions, and techniques for resolving interactions at run-time. While this pioneering work was foundational and very successful, it is limited in the sense that it is based on assumptions that hold only for telecommunication systems. In the meantime, different notions of feature interactions have emerged in different communities, including Internet applications, service systems, adaptive systems, automotive systems, software product lines, requirements engineering, and computational biology. So, feature interactions are a much more general concept than investigated in the past in the context of telecommunication systems, but a classification, comparison, and generalization of the multitude of different views is missing. The feature-interaction problem is still of pivotal importance in various industrial applications, and the Dagstuhl seminar “Feature Interactions: The Next Generation” gathered researchers and practitioners from different areas of computer science and other disciplines with the goal to compare, discuss, and consolidate their views, experience, and domain-specific solutions to the feature-interaction problem.

Seminar July 6–11, 2014 – <http://www.dagstuhl.de/14281>

1998 ACM Subject Classification D.2.1 Requirements/Specifications, D.2.4 Software/Program Verification, D.2.10 Design, D.2.11 Software Architectures, D.2.13 Reusable Software

Keywords and phrases Feature interactions, feature-interaction problem, feature orientation, product lines, modularity, composition

Digital Object Identifier 10.4230/DagRep.4.7.1

Edited in cooperation with Sergiy Kolesnikov



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

Feature Interactions: The Next Generation, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 1–24

Editors: Sven Apel, Joanne M. Atlee, Luciano Baresi, and Pamela Zave



DAGSTUHL
REPORTS

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

1 Executive Summary

Sven Apel

Joanne M. Atlee

Luciano Baresi

Pamela Zave

License © Creative Commons BY 3.0 Unported license
© Sven Apel, Joanne M. Atlee, Luciano Baresi, and Pamela Zave

Overview and Motivation

A major goal of software and systems engineering is to construct systems from reusable parts, which we call *features* (end-user-visible units of behavior or increments in system functionality). Such a compositional approach can decrease time to market, improve product quality, and diversify the product portfolio. However, the success of a compositional approach depends on the modularity of the reusable parts. The quest for modularity has a long tradition in software and systems engineering, programming languages research, and even in newer fields such as synthetic biology.

In the early days of software and systems engineering, the feature-interaction problem was identified (and coined) as a major threat to modularity [8, 31, 25]. A feature interaction occurs when the behavior of one feature is affected by the presence of another feature. Often the interaction cannot be deduced easily from the intended behaviors of the individual features involved. A canonical example is the inadvertent interaction between the call-forwarding and call-waiting features of a telephony system [8]: If both features are active, the system can reach an undefined possibly unsafe state when it receives a call on a busy line, because it is not specified whether the call should be suspended or forwarded. Alternatively, a feature interaction can be *planned*: for example, advanced cruise-control features are designed to interact with and extend basic cruise control.

To be safe, software developers must analyze the consequences of all possible feature interactions, in order to find and fix the undesired interactions. The *feature-interaction problem* is that the number of potential interactions to consider is exponential in the number of features. As a result, software developers find that their work in developing new features is dominated by the tasks to detect, analyze, and verify interactions.

The feature-interaction problem is deeply rooted in the fact that the world is often not compositional [25, 20]. That is, a feature is not an island. It communicates and cooperates with other features and the environment, so it cannot be completely isolated. Insights from complex-systems research suggest that feature interactions are a form of emergent behavior that is inherent to any system that consists of many, mutually interacting parts. So, emergent system behavior – which is not deducible from the individual parts of a system – can be observed in many situations including in quantum systems (e. g., superconductivity), biological systems (e. g., swarm intelligence), and economical systems (e. g., trading market crashes). The challenge is to foster and manage *desired* interactions and to detect, resolve, and even avoid *undesired* feature interactions – in a scalable manner.

The feature-interaction problem became a crisis in the telecommunications industry in the late 1980s [5]. To handle complexity, there was the strong desire to *compose* systems from independently developed features, but there was no means to detect, express, and reason about feature interactions. Researchers responded with formalisms that enable automatic detection of feature interactions [4, 7, 15, 14, 21, 26], architectures that avoid classes of interactions [17, 29, 18, 28, 31], and techniques for resolving interactions at run-

time [16, 27]. Architectural solutions have been the most successful because they impose general coordination strategies (i. e., serial execution) that apply to all features that are ‘plugged’ into the architecture, thereby, addressing the scalability issue at the heart of the feature-interaction problem. In coordination-based approaches, such as BIP [2, 3] or Composition Patterns [10], the interactions among a set of features are specified explicitly and can be specialized for subsets of features.

While the pioneering work on the feature-interaction problem in telecommunication systems was foundational and very successful [8], it is limited in the sense that it is based on assumptions that hold for telecommunication systems, but that do not hold in other domains. For example, architecture-based approaches take advantage of the fact that communication takes place over a mostly serial connection between communicating parties – which is not the case in systems made up of parallel components (e. g., service systems, automotive software) or software product lines (e. g., features implemented via conditional compilation such as the Linux kernel). Specifying interactions explicitly is not a general solution either. When facing systems composed of thousands of features, attempting to identify and model a possibly exponential number of feature interactions is elusive. Furthermore, the highly dynamic nature of feature (or service) composition in self-adaptive systems, dynamic product lines, cloud computing, and systems of systems imposes a new class of challenges to solving the feature-interaction problem [24, 9, 1].

So, it is not surprising that different notions of feature interactions have emerged in different communities [6]. Instances of the feature-interaction problem have been observed and addressed in Internet applications [11], service systems [30], automotive systems [12], software product lines [19], requirements engineering [23], computational biology [13], and in many other fields outside of computer science. While all instances of the problem are rooted in the nature of modularity and compositionality [25, 20], the individual views, interpretations, and possible solutions differ considerably. For example, the view on feature interactions taken in program synthesis [22] differs significantly from the view in automotive systems engineering [12]: there are structural vs. behaviour views, static vs. dynamic views, sequential vs. parallel views, functional vs. non-functional, coordinated vs. emergent-behaviour views, and so on. It turns out that feature interactions are a much more general concept than investigated in the past in the context of telecommunication systems, but a classification, comparison, and generalization of the multitude of different views is missing.

The feature-interaction problem is still of pivotal importance in various industrial applications, but, despite significant efforts, it is far from being solved. The underlying hypothesis of organizing a Dagstuhl seminar on this topic was that the time is ripe to gather researchers and practitioners from different areas of computer science and other disciplines to compare, discuss, and consolidate their views, experience, and domain-specific solutions to the feature-interaction problem. To make progress, scientific discourse on the feature-interaction problem must be based on a broader foundation to be able to join forces of different communities. Can other domains learn from the success of domain-specific solutions for telecommunication systems? Are there key principles, patterns, and strategies to represent, identify, manage, and resolve feature interactions that are domain-independent, that are valid and useful across domains? Or, should we strive for domain-specific solutions that are only loosely related to solutions from other domains? Can we develop a unified terminological and conceptual framework for feature-interaction research? Is that even possible or meaningful, given that interactions in telecommunication systems and emergent behavior and phase transitions in swarm systems are, although related, quite different views?

Goals of the Seminar and Further Activities

It is our goal and firm belief that the feature-interaction problem needs to be viewed from a broader perspective. While feature interactions are still a major challenge in software and systems engineering, both in academia and industry, research on the feature-interaction problem has diversified and diverged in the last decade. Researchers working on similar problems, but in different contexts, are largely disconnected and unaware of related work. A major goal of the seminar was to (re)launch a sustained research community that embraces researchers and practitioners from different fields within and outside computer science. We firmly believe that we reached this goal with our seminar. In particular, a subset of the participants is going to organize a follow-up seminar that directly builds on this seminar's results. The next major milestone will be – now as we gained a better understanding of the similarities and differences between the different notions of feature interactions – to establish a catalog on feature-interaction patterns and solutions thereof. The idea for this pattern catalog arose from the final panel session of the seminar. It is inspired by work on patterns in architecture (of buildings). Such a catalog will be the necessary basis for further research on leveraging patterns for detecting, managing, and resolving feature interactions in different kinds of systems.

References

- 1 L. Baresi, S. Guinea, and L. Pasquale. Service-oriented dynamic software product lines. *IEEE Computer*, 45(10):42–48, 2012.
- 2 A. Basu, M. Bozga, and J. Sifakis. Modeling heterogeneous real-time components in BIP. In *Proc. of the Int'l Conf. on Software Engineering and Formal Methods (SEFM)*, pages 3–12. IEEE, 2006.
- 3 S. Bliudze and J. Sifakis. The algebra of connectors – Structuring interaction in BIP. *IEEE Transactions on Computers*, 57(10):1315–1330, 2008.
- 4 J. Blom, B. Jonsson, and L. Kempe. Using temporal logic for modular specification of telephone services. In *Feature Interactions in Telecommunications Systems*, pages 197–216. IOS Press, 1994.
- 5 T. Bowen, F. Dworack, C. Chow, N. Griffeth, G. Herman, and Y.-J. Lin. The feature interaction problem in telecommunications systems. In *Proc. of the Int'l Conf. on Software Engineering for Telecommunication Switching Systems (SETSS)*, pages 59–62. IEEE, 1989.
- 6 G. Bruns. Foundations for features. In *Feature Interactions in Telecommunications and Software Systems VIII*, pages 3–11. IOS Press, 2005.
- 7 G. Bruns, P. Mataga, and I. Sutherland. Features as service transformers. In *Feature Interactions in Telecommunications Systems V*, pages 85–97. IOS Press, 1998.
- 8 M. Calder, M. Kolberg, E. Magill, and S. Reiff-Marganiec. Feature interaction: A critical review and considered forecast. *Computer Networks*, 41(1):115–141, 2003.
- 9 B. Cheng, R de Lemos, H. Giese, P. Inverardi, J. Magee, et al. Software engineering for self-adaptive systems: A research roadmap. In *Software Engineering for Self-Adaptive Systems*, LNCS 5525, pages 1–26. Springer, 2009.
- 10 S. Clarke and R. Walker. Composition patterns: An approach to designing reusable aspects. In *Proc. of the Int'l Conf. on Software Engineering (ICSE)*, pages 5–14. IEEE, 2001.
- 11 R. Crespo, M. Carvalho, and L. Logrippo. Distributed resolution of feature interactions for Internet applications. *Computer Networks*, 51(2):382–397, 2007.
- 12 A. Dominguez. *Detection of Feature Interactions in Automotive Active Safety Features*. PhD thesis, School of Computer Science, University of Waterloo, 2012.
- 13 R. Donaldson and M. Calder. Modular modelling of signalling pathways and their cross-talk. *Theoretical Computer Science*, 456(0):30–50, 2012.

- 14 A. Felty and K. Namjoshi. Feature specification and automated conflict detection. *ACM Transactions on Software Engineering and Methodology*, 12(1):3–27, 2003.
- 15 M. Frappier, A. Mili, and J. Desharnais. Defining and detecting feature interactions. In *Proc. of the IFIP TC 2 WG 2.1 Int'l Workshop on Algorithmic Languages and Calculi*, pages 212–239. Chapman & Hall, Ltd., 1997.
- 16 N. Griffeth and H. Velthuisen. The negotiating agents approach to runtime feature interaction resolution. In *Feature Interactions in Telecommunications Systems*, pages 217–235. IOS Press, 1994.
- 17 J. Hay and J. Atlee. Composing features and resolving interactions. In *Proc. of the ACM SIGSOFT Symp. on Foundations of Software Engineering (FSE)*, pages 110–119. ACM, 2000.
- 18 M. Jackson and P. Zave. Distributed feature composition: A virtual architecture for telecommunications services. *IEEE Transactions on Software Engineering (TSE)*, 24(10):831–847, 1998.
- 19 P. Jayaraman, J. Whittle, A. Elkhodary, and H. Gomaa. Model composition in product lines and feature interaction detection using critical pair analysis. In *Proc. of the Int'l Conf. on Model Driven Engineering Languages and Systems (MoDELS)*, LNCS 4735, pages 151–165. Springer, 2007.
- 20 C. Kästner, S. Apel, and K. Ostermann. The road to feature modularity? In *Proc. of the Int'l Workshop on Feature-Oriented Software Development (FOSD)*, pages 5:1–5:8. ACM, 2011.
- 21 F. Lin and Y.-J. Lin. A building block approach to detecting and resolving feature interactions. In *Feature Interactions in Telecommunications Systems*, pages 86–119. IOS Press, 1994.
- 22 J. Liu, D. Batory, and C. Lengauer. Feature oriented refactoring of legacy applications. In *Proc. of the Int'l Conf. on Software Engineering*, pages 112–121. ACM, 2006.
- 23 A. Nhlabatsi, R. Laney, and B. Nuseibeh. Feature interaction: The security threat from within software systems. *Progress in Informatics*, (5):75–89, 2008.
- 24 L. Northrop, P. Feiler, R. Gabriel, J. Goodenough, R. Linger, T. Longstaff, R. Kazman, M. Klein, D. Schmidt, K. Sullivan, and K. Wallnau. Ultra-large-scale systems – The software challenge of the future. Technical report, Software Engineering Institute, Carnegie Mellon University, 2006.
- 25 K. Ostermann, P. Giarrusso, C. Kästner, and T. Rendel. Revisiting information hiding: Reflections on classical and nonclassical modularity. In *Proc. of the Europ. Conf. on Object-Oriented Programming (ECOOP)*, LNCS 6813, pages 155–178, 2011.
- 26 K. Pomakis and J. Atlee. Reachability analysis of feature interactions: A progress report. In *Proc. of the Int'l Symp. on Software Testing and Analysis (ISSTA)*, pages 216–223. ACM, 1996.
- 27 S. Tsang and E. Magill. Learning to detect and avoid run-time feature interactions in intelligent networks. *IEEE Transactions on Software Engineering (TSE)*, 24(10):818–830, 1998.
- 28 G. Utas. A pattern language of feature interaction. In *Feature Interactions in Telecommunications Systems V*, pages 98–114. IOS Press, 1998.
- 29 R. van der Linden. Using an architecture to help beat feature interaction. In *Feature Interactions in Telecommunications Systems*, pages 24–35. IOS Press, 1994.
- 30 M. Weiss, B. Esfandiari, and Y. Luo. Towards a classification of web service feature interactions. *Computer Networks*, 51(2):359–381, 2007.
- 31 Pamela Zave. Modularity in Distributed Feature Composition. In *Software Requirements and Design: The Work of Michael Jackson*, pages 267–290. Good Friends Publishing, 2010.

2 Table of Contents

Executive Summary

<i>Sven Apel, Joanne M. Atlee, Luciano Baresi, and Pamela Zave</i>	2
--	---

Perspective Talks

Toward User-Centric Feature Composition for the Internet of Things <i>Pamela Zave</i>	8
The Feature Interaction Problem in a Federated Communications-Enabled Collaboration Platform <i>Mario Kolberg</i>	8
Feature Interactions in Software Systems: An Implementation Perspective <i>Christian Kästner, Sven Apel</i>	9
Feature Interactions in Smartphones <i>Christian Prehofer</i>	10
Behaviours and Feature Interactions <i>Michael Jackson</i>	11

Lightning Talks

Extracting Feature Model Changes from the Linux Kernel using FMDiff <i>Nicolas Dintzner</i>	12
Feature Interactions Taxonomy and Case Studies <i>Sergiy Kolesnikov</i>	12
(Structural) Feature Interactions for Variability-Intensive Systems Testing <i>Gilles Perrouin</i>	13
Performance Prediction in the Presence of Feature Interactions <i>Norbert Siegmund</i>	13
Feature Interaction in the Browser and the Software-Defined Network <i>Shriram Krishnamurthi</i>	15
Feature Interaction and Emergent Properties <i>Gerhard Chroust</i>	15
Extending Ruby into a DSL Good and Bad Feature Interactions <i>Thomas Gschwind</i>	16
Probabilistic Model Checking of DTMC Models of User Activity Patterns <i>Oana M. Andrei</i>	17
Presence-Condition Simplification <i>Alexander von Rhein</i>	17
On the Relation between Feature Dependencies and Change Propagation <i>Bruno Cafeo</i>	18

Breakout Groups: Domain-independence of Feature Interactions

Summary of Group 1 <i>Krzysztof Czarnecki</i>	18
--	----

Summary of Group 2	
<i>Sandro Schulze, Sebastian Erdweg</i>	19
Summary of Group 3	
<i>Oscar M. Nierstrasz</i>	20
Breakout Groups: Framework for Modeling Feature Interactions	
Summary of Group 1	
<i>Michael Jackson</i>	21
Summary of Group 2	
<i>Kathi Fisler</i>	22
Panel Discussions	
Reflections and Perspectives	
<i>Sven Apel</i>	22
Participants	24

3 Perspective Talks

3.1 Toward User-Centric Feature Composition for the Internet of Things

Pamela Zave (AT&T Labs Research, US)

License © Creative Commons BY 3.0 Unported license
© Pamela Zave

Main reference P. Zave, E. Cheung, S. Yarosh, “Toward User-Centric Feature Composition for the Internet of Things,” unpublished manuscript.

URL <http://www2.research.att.com/~pamela/userFtrComp.pdf>

Many user studies of home automation, as the most familiar representative of the Internet of Things, have shown the difficulty of developing technology that users understand and like. It helps to state requirements as largely-independent features, but features are not truly independent, so this incurs the cost of managing and explaining feature interactions. We propose to compose features at runtime, resolving their interactions by means of priority. Although the basic idea is simple, its details must be designed to make users comfortable by balancing manual and automatic control. On the technical side, its details must be designed to allow meaningful separation of features and maximum generality. As evidence that our composition mechanism achieves its goals, we present three substantive examples of home automation, and the results of a user study to investigate comprehension of feature interactions. A survey of related work shows that this proposal occupies a sensible place in a design space whose dimensions include actuator type, detection versus resolution strategies, and modularity.

3.2 The Feature Interaction Problem in a Federated Communications-Enabled Collaboration Platform

Mario Kolberg (University of Stirling, Stirling, Scotland, GB)

License © Creative Commons BY 3.0 Unported license
© Mario Kolberg

Joint work of Kolberg, M.; Buford, J. F.; Dhara, K.; Wu, X.; Krishnaswamy, V.

Main reference M. Kolberg, J. F. Buford, K. Dhara, X. Wu, V. Krishnaswamy, “Feature Interaction in a Federated Communications-Enabled Collaboration Platform,” *Computer Networks Journal*, 57(12):2410–2428, 2013.

URL <http://dx.doi.org/10.1016/j.comnet.2013.02.023>

For enterprise use there is a need to integrate various collaboration tools such as email, instant messages, wikis, blogs, web conferences, and shared documents, as well as link with existing intelligent communication systems to support long-term collaborations in a variety of ways. By the very nature of such systems they include a large number of independently developed features and services and thus provide a strong potential for feature interactions. This paper presents novel work on feature interaction analysis in collaboration environments and presents new types of interactions found in this space.

In this talk ConnectedSpaces is used as a basis to carry out a detailed analysis of feature interaction problems in collaboration environments. ConnectedSpaces is a new model for federated collaboration environments. Like a number of existing systems, ConnectedSpaces uses a collaboration space as the basic construct. ConnectedSpaces enables the user to work directly in the client application of their choice, including MS Outlook, Internet Explorer and Skype.

This talk presents distinctive characteristics of ConnectedSpaces, including views, spaces as communication endpoints, space persistence and structuring, and embedded objects. Using these features, new types of feature interactions for collaboration platforms are categorized and analyzed. This work is novel as it is the first investigation into feature interactions with collaboration platforms. The talk will also outline potential approaches to handle such interactions. We advocate a runtime feature interaction technique which can cope with features being provided by different organizations.

3.3 Feature Interactions in Software Systems: An Implementation Perspective

Christian Kästner (Carnegie Mellon University, US), Sven Apel (University of Passau, DE)

License © Creative Commons BY 3.0 Unported license

© Christian Kästner, Sven Apel

Joint work of Apel, Sven; Kästner, Christian; Nguyen, Hung Viet; Nguyen, Tien N.; Kolesnikov, Sergiy; Siegmund, Norbert

Main reference S. Apel, S. Kolesnikov, N. Siegmund, C. Kästner, B. Garvin, “Exploring Feature Interactions in the Wild: The New Feature-Interaction Challenge,” in Proc. of the 5th Int’l Workshop on Feature-Oriented Software Development (FOSD’13), pp. 1–8, ACM, 2013.

URL <http://dx.doi.org/10.1145/2528265.2528267>

In this talk, we have discussed feature interactions in software systems from an implementation perspective. Feature interactions are a common problem in configurable systems, among implementations of configuration options. Key differences to classic research on feature interactions [3] are: We operate under a closed-world assumption where the implementation of all features are known; we typically focus on a single, non-distributed process; variability is induced by configuration options with non-trivial dependencies, implemented through various implementation mechanisms, from modules to conditional compilation. Feature interactions manifest in different forms, four of which we discussed.

First, interaction bugs occur when a system exhibits a bug if and only if multiple options are selected in specific combinations [5]. Typically options work well in isolation, but expose a bug if combined. The community has developed close-world, whole-product-line techniques, which we call variability-aware or family-based analyses [1, 7], that can identify certain classes of bugs. Compared to standard analysis techniques, variability-aware analyses cover the whole configuration space and allow statements about the configurable system in its entirety (and not only about individual configurations). This way, several bugs have been found that are only triggered by specific configurations settings. Empirical studies on the reported bugs have revealed common interaction patterns.

Second, performance interactions have been defined as unexpected performance behaviors when combining multiple configuration options [6]. Assuming that the influence of each option on the system’s performance (for a given benchmark) can be isolated, a performance interaction occurs (according to our definition) when the performance of multiple options cannot be explained by their individual performances. Automated interaction detection based on sampling techniques has been successful in improving performance prediction in configurable systems; it has found that performance interactions occur mostly among pairs of options, but also beyond.

Third, at the level of source code, interactions are manifested as glue code that combines or coordinates the implementation of multiple options; the additional code is only included in the program if all options are selected. This pattern is common in component connectors,

lifters, derivatives, connector plugins, as well as code in nested `ifdef` directives and `if` statements. Code-level interactions are relatively easy to identify (especially, for compile-time configuration mechanisms), but they are not necessarily useful predictors of other kinds of interactions.

Finally, in a running system, we consider it as an interaction when the value of a program variable depends on multiple configuration options [4]. For example, when executing a test case in a configurable system, we expect most variables to have the same value in all executed configurations, and a few variables to have two (or more) alternative values depending on a single option. However, variables can potentially depend on many options. In a study of Wordpress, we found dependencies among up to 16 of 50 optional plugins.

Overall, the picture of feature interactions in configurable systems is diverse, and there is no single, feasible classification or terminology. Working with real systems provides lots of data and much opportunity for studying interactions. While some of these interactions are easy and reliable to detect, others require intensive testing and sophisticated sampling. We hope that, in the long run, we are able to identify correlations between different kinds of interactions, found with different techniques, and to combine them effectively [2].

References

- 1 S. Apel, D. Batory, C. Kästner, and G. Saake. *Feature-Oriented Software Product Lines: Concepts and Implementation*. Springer, October 2013.
- 2 S. Apel, S. Kolesnikov, N. Siegmund, C. Kästner, and B. Garvin. Exploring Feature Interactions in the Wild: The New Feature-Interaction Challenge. In *Proc. of the Int'l Workshop on Feature-Oriented Software Development (FOSD)*, pages 1–8. ACM, 2013.
- 3 M. Calder, M. Kolberg, E. Magill, and S. Reiff-Marganiec. Feature Interaction: A Critical Review and Considered Forecast. *Computer Networks*, 41(1):115–141, 2003.
- 4 H. Nguyen, C. Kästner, and T. Nguyen. Exploring Variability-Aware Execution for Testing Plugin-Based Web Applications. In *Proc. of the Int'l Conf. on Software Engineering (ICSE)*, pages 907–918. ACM, 2014.
- 5 C. Nie and H. Leung. A Survey of Combinatorial Testing. *ACM Computing Surveys*, 43(2):1–29, 2011.
- 6 N. Siegmund, S. Kolesnikov, C. Kästner, S. Apel, D. Batory, M. Rosenmüller, and G. Saake. Predicting Performance via Automated Feature-Interaction Detection. In *Proc. of the Int'l Conf. on Software Engineering (ICSE)*, pages 167–177. IEEE, 2012.
- 7 T. Thüm, S. Apel, C. Kästner, I. Schaefer, and G. Saake. A Classification and Survey of Analysis Strategies for Software Product Lines. *ACM Computing Surveys*, 47(1):6:1–6:45, 2014.

3.4 Feature Interactions in Smartphones

Christian Prehofer (fortiss GmbH – München, DE)

License  Creative Commons BY 3.0 Unported license
© Christian Prehofer

This talk reviews feature interactions in telecom, mobile phones and smartphones. We first discuss the history of feature interactions in telecommunications and focus on typical reasons for feature interactions. In particular, Software and standards evolution over long period of time as well as legacy devices are a frequent reason behind feature interactions. Then, we discuss technology developments in telephony systems, standards as well as system platforms which relate to these causes of feature interactions.

3.5 Behaviours and Feature Interactions

Michael Jackson (The Open University – Milton Keynes, GB)

License  Creative Commons BY 3.0 Unported license
© Michael Jackson

- The feature interaction problem arises in the physical problem world of computer-based systems rather than in the machine – the software – per se. The difficulty, in essence, is that each feature places its own demands on the physical behaviour of the problem world, and that the demands of different features may be in some way incompatible. In automotive software, the speed limiting and cruise control features may conflict: the car can have only one speed at any one time. In controlling an elevator, normal use conflicts with use by firefighters: when the lift stops at a floor, normal use demands that the doors close automatically after a specified delay, but firefighter use demands that they close only in response to the Door_Close button in the lift car.
- Direct conflict is far from the only type of feature interaction. Additional interaction types include: mutual exclusion, interference, resource sharing, and – very commonly – switching, in which control of some part of the problem world is passed from one regime that has terminated to another that has been newly activated. Further, the physical nature of the problem world can vitiate apparently sound reasoning: the fact that each of two physical demands can be satisfied in isolation gives no guarantee that they can be satisfied in combination. Even when the alphabets of the two demands appear to be disjoint they may interact through the medium of other phenomena that have been neglected in the analysis.
- In computer-based systems, which interact with and control the physical world, our primary concerns are with system dynamics. The behaviour of the system – of the interacting computing machine and problem world – is the essential product of software development, and the salient aspect of this behaviour is in the problem world, not in the software. In developing the software we are developing this behaviour, and we may usefully regard feature interaction as the interaction among the constituent behaviours that together make up the whole behaviour of the system. Because the behaviour of a realistic – and especially a critical – system is very complex, it is necessary to adopt a disciplined approach to its design.
- The design approach suggested here structures the system behaviour into its constituent behaviours by a combination of top-down and bottom-up decomposition. Each constituent behaviour is considered in its totality: that is, the desired problem world behaviour is considered together with the software behaviour that will evoke it and also the behaviours of all parts of the problem world that lie implicitly on a causal path between the machine and the problem world behaviour explicitly desired.
- Further elements of the suggested approach are important. First, each proposed constituent behaviour is initially considered in a loose decomposition: it is considered in isolation, as if it were a complete stand-alone system in itself, ignoring its eventual interactions with other constituent behaviours. Second, the complexity of each constituent behaviour is developed in stages: the main-line isolated behaviour; the main-line behaviour elaborated as necessary to handle exceptional conditions; and the elaborated behaviour further complicated by its interactions with other constituent behaviours. Third, the control of behaviour is rigorously separated from behaviour content.
- This approach offers advantages in the identification and treatment of feature interactions. Considering each behaviour in its totality maintains an awareness of the physical implica-

tions of the design both of the desired behaviour and of the software behaviour that will evoke it. This awareness is strengthened by the emphasis on simplicity, separating out the sources of behavioural complexity. The insistence on loose decomposition ensures that combination of the constituent behaviours to give the complete system behaviour is a distinct and explicitly recognised design task in which the constituents to be combined are already well understood. In this combination task, feature interactions are the heart of the design problem to be addressed, and much – perhaps, everything possible – has been done to make the task and its design problem as perspicuous as it can be.

4 Lightning Talks

4.1 Extracting Feature Model Changes from the Linux Kernel using FMDiff

Nicolas Dintzner (TU Delft, NL)

License © Creative Commons BY 3.0 Unported license
© Nicolas Dintzner

Joint work of Dintzner, Nicolas; Van Deursen, Arie; Pinzger, Martin

Main reference N. Dintzner, A. Van Deursen, M. Pinzger, “Extracting feature model changes from the Linux kernel using FMDiff,” in Proc. of the 8th Int’l Workshop on Variability Modelling of Software-Intensive Systems (VaMoS’14), Article No. 22, ACM, 2014.

URL <http://dx.doi.org/10.1145/2556624.2556631>

The Linux kernel feature model has been studied as an example of large scale evolving feature model and yet details of its evolution are not known. We present here a classification of feature changes occurring on the Linux kernel feature model, as well as a tool, FMDiff, designed to automatically extract those changes. With this tool, we obtained the history of more than twenty architecture specific feature models, over ten releases and compared the recovered information with Kconfig file changes. We establish that FMDiff provides a comprehensive view of feature changes and show that the collected data contains promising information regarding the Linux feature model evolution.

4.2 Feature Interactions Taxonomy and Case Studies

Sergiy Kolesnikov (University of Passau, DE)

License © Creative Commons BY 3.0 Unported license
© Sergiy Kolesnikov

Joint work of Apel, Sven; Kolesnikov, Sergiy; Siegmund, Norbert; Kästner, Christian; Garvin, Brady

Main reference S. Apel, S. Kolesnikov, N. Siegmund, C. Kästner, B. Garvin, “Exploring Feature Interactions in the Wild: The New Feature-Interaction Challenge,” in Proc. of the 5th Int’l Workshop on Feature-Oriented Software Development (FOSD’13), pp. 1–8, ACM, 2013.

URL <http://dx.doi.org/10.1145/2528265.2528267>

The *feature-interaction problem* has been keeping researchers and practitioners in suspense for years. Although there has been substantial progress in developing approaches for modeling, detecting, managing, and resolving feature interactions, we lack sufficient knowledge on the kind of feature interactions that occur in real-world systems. In this talk, we set out the goal to explore the nature of feature interactions systematically and comprehensively, classified in terms of order and visibility. Understanding this nature will have significant implications on research in this area, for example, on the efficiency of interaction-detection

or performance-prediction techniques. A set of preliminary results as well as a discussion of possible experimental setups and corresponding challenges give us confidence that this endeavor is within reach but requires a collaborative effort of the community.

4.3 (Structural) Feature Interactions for Variability-Intensive Systems Testing

Gilles Perrouin (University of Namur, BE)

License © Creative Commons BY 3.0 Unported license
© Gilles Perrouin

Joint work of Perrouin, Gilles; Henard, Christopher; Papadakis, Mike; Klein, Jacques; Heymans, Patrick; Le Traon, Yves

Main reference C. Henard, M. Papadakis, G. Perrouin, J. Klein, P. Heymans, Y.L. Traon, “Bypassing the combinatorial explosion: Using similarity to generate and prioritize t-wise test configurations for software product lines,” *IEEE Transactions on Software Engineering*, 40(7):650–670, 2014.

URL <http://dx.doi.org/10.1109/TSE.2014.2327020>

Adopting a middle-ground view between Michael Jackson’s (“features are behaviours”) and Christian Kaestner’s (“features are configuration options”) definitions, we consider features as units of variability specified within a feature model. Feature models can be used to document valid choices (called configurations) formed by combinations of features. Such configurations can either relate to desired elevator behaviours interactions (normal use, emergency use, etc.) or to viable Linux kernels. The number of configurations derivable from a given feature model grows exponentially with the number of features, making the testing process inherently difficult. To harness combinatorial explosion of the number of configurations to be considered, we propose to sample them by computing t-way interactions from the feature model. We present initial experiments and a search-based approach maximising dissimilarity between configurations. This approach mimics combinatorial interaction testing techniques in a flexible and scalable manner.

References

- 1 Gilles Perrouin, Sebastian Oster, Sagar Sen, Jacques Klein, Benoit Baudry, and Yves le Traon. Pairwise testing for software product lines: comparison of two approaches. *Software Quality Journal*, 20(3-4):605–643, 2012.
- 2 Christopher Henard, Mike Papadakis, Gilles Perrouin, Jacques Klein, and Yves Le Traon. Pledge: A product line editor and test generation tool. In *Proceedings of the 17th International Software Product Line Conference Co-located Workshops, SPLC’13 Workshops*, pages 126–129, New York, NY, USA, 2013. ACM.

4.4 Performance Prediction in the Presence of Feature Interactions

Norbert Siegmund (University of Passau, DE)

License © Creative Commons BY 3.0 Unported license
© Norbert Siegmund

Joint work of Siegmund, Norbert; Kolesnikov, Sergiy; Christian, Kästner; Apel, Sven; Batory, Don; Rosenmüller, Marko; Saake, Gunter

Customizable programs and program families provide user-selectable features allowing users to tailor the programs to the application scenario. Beside functional requirements, users

are often interested in non-functional requirements, such as a binary-size limit, a minimized energy consumption, and a maximum response time.

In our work, we aim at predicting a configuration's non-functional properties for a specific workload based on the user-selected features [2, 3, 4]. To this end, we quantify the influence of each selected feature on a non-functional property to compute the properties of a specific configuration. Here, we concentrate on performance only.

Unfortunately, the accuracy of performance predictions may be low when considering features only in isolation, because many factors influence performance. Usually, a property is program-wide: it emerges from the presence and interplay of multiple features. For example, database performance depends on whether a search index or encryption is used and how both features interplay. If we knew how the combined presence of two features influences performance, we could predict a configuration's performance more accurately. Two features interact (i. e., cause a performance interaction) if their simultaneous presence in a configuration leads to an unexpected performance, whereas their individual presences do not. We improve the accuracy of predictions in two steps: (i) We detect which features interact and (ii) we measure to what extent they interact. In our approach, we aim at finding the sweet spot between prediction accuracy, measurement effort, and generality in terms of being independent of the application domain and the implementation technique. The distinguishing property of our approach is that we neither require domain knowledge, source code, nor complex program-analysis methods, and our approach is not limited to special implementation techniques, programming languages, or domains.

Our key idea to determine which features interact is the following: We measure each feature twice. In the first run, we try to measure the performance influence of the feature in isolation by measuring the variant that has the smallest number of additionally selected features. The second run, aims at maximizing the number of features such that all possible interactions that may influence on performance materialize in the measurement. If the influence of the feature in isolation differs with the influence when combined with other features, we know that this feature interacts. In the second step, we perform several sampling heuristics, such as pair-wise sampling, to determine the actual combinations of interacting features that cause interactions.

Our evaluation is based on six real-world case studies from varying domains (e. g., databases, encoding libraries, and web servers) using different configuration techniques. Our experiments show an average prediction accuracy of 95 percent, which is a 15 percent improvement over an approach that takes no interactions into account [1].

References

- 1 N. Siegmund, S. Kolesnikov, C. Kästner, S. Apel, D. Batory, M. Rosenmüller, and G. Saake. Predicting Performance via Automated Feature-Interaction Detection. In *Proc. ICSE*, pages 167–177. IEEE, 2012.
- 2 N. Siegmund, M. Rosenmüller, C. Kästner, P. Giarrusso, S. Apel, and S. Kolesnikov. Scalable Prediction of Non-functional Properties in Software Product Lines. In *Proc. SPLC*, pages 160–169. IEEE, 2011.
- 3 N. Siegmund, M. Rosenmüller, C. Kästner, P. Giarrusso, S. Apel, and S. Kolesnikov. Scalable Prediction of Non-functional Properties in Software Product Lines: Footprint and Memory Consumption. *Information and Software Technology*, 55(3):491–507, 2013.
- 4 Norbert Siegmund, Alexander von Rhein, and Sven Apel. Family-Based Performance Measurement. In *Proceedings of the International Conference on Generative Programming and Component Engineering (GPCE)*, pages 95–104. ACM, 2013.

4.5 Feature Interaction in the Browser and the Software-Defined Network

Shriram Krishnamurthi (Brown University, US)

License © Creative Commons BY 3.0 Unported license
© Shriram Krishnamurthi

In this presentation, I give a brief overview of feature interaction problems as they occur in current real-world systems and are likely to occur in future ones. I illustrate the present using Web browser extensions and the problem of finding violations (especially of security-sensitive properties such as Private Browsing Mode). For the future, I speculate that software-defined networking will result in “app stores” of networking behavior, which will inevitably interact in unsavory ways and will need to be kept distinct.

4.6 Feature Interaction and Emergent Properties

Gerhard Chroust (Johannes Kepler Universität Linz, AT)

License © Creative Commons BY 3.0 Unported license
© Gerhard Chroust

Main reference G. Chroust, “System properties under composition,” in Proc. of the European Meeting on Cybernetics and Systems Research (EMCSR’02), pp. 203–208, Austrian Society for Cybernetic Studies, 2002.

The interaction of features is a mixed blessing in engineering. It enables useful functions e. g. radio reception by utilizing resonance, but it can also have disastrous and destructive effects like the collapse of a bridge due to undesirable resonance. In systems theory the so-called emergent properties of systems with individual components show similar effects. In this paper I explore the similarities between Feature Interaction in Systems Engineering and Emergent Properties in Systems Theory and show the analogy between these two concepts.

According to my understanding (following K.C. Kang, 1990) “*a feature of a . . . product can be described as a prominent or distinctive user-visible aspect, quality, or characteristic . . .*”.

When composing a system from interconnected subsystems (components) the properties of the resulting system (and their predictability!) are one of the key issues of engineering. Difficulties stem from two observations:

- The structure of the system plays a key role.
- A composed system often exhibits ‘unexpected’ behavior due to the occurrence of so-called emergent properties.

In general a system’s properties depend on the structure of the system and on all properties of all components in the system AND are usually different from the properties of the individual components.

We define an “*emergent property of a system is a property which cannot be determined solely from the properties of the system’s components, but which is additionally determined by the system’s structure (i. e. by the way the parts are connected to form the system)*”.

Emergent properties depend inherently and essentially on the structure of the system i. e. on the way the system is composed, and change with a change of the structure! It is often not easy to determine them from the system, they often appear as a surprise (“emergence”).

It is obvious that emergent properties complicate the prediction of system properties since their value can only be determined by considering also the system structure. As

long as we allow any imaginable structure for a system there is little chance to make any reasonable statements about the emergent properties. A restricted set of admissible composition structures allows to make some statements about the behavior of emergent properties under composition. Software patterns are good candidates for such “admissible composition patterns.”

All Features are system properties, but not all system properties are to be considered Features. Feature Interactions can be interpreted as emergent properties since they usually depend on the (static and/or dynamic) structure of the system.

One of the standard examples in the seminar was an automatic door locking/unlocking mechanism which was dependent on the time-of-the-day, the setting of several switches with different purposes and a timeout facility. The resulting feature could be called “safety of the house” and is a typical “emergent property,” where both the physical structure (whether some components are arranged in parallel or serial order) and also the chronological order (in which certain settings are activated) have a decisive influence on the outcome.

My contribution established the strong analogy between Feature Interaction in systems engineering and the system theoretic concept of emergent properties in multi-component systems. This analogy can be used to consolidate the terminology and exchange insights between these two domains. I hope that this will be a source for fruitful discussion and more clarification in both areas.

4.7 Extending Ruby into a DSL Good and Bad Feature Interactions

Thomas Gschwind (IBM Research GmbH – Zürich, CH)

License © Creative Commons BY 3.0 Unported license
© Thomas Gschwind

Joint work of Michael H. Kalantar, James Doran, Tamar Eilam, Michael D. Elder, Fabio Oliveira, Edward C. Snible, Tova Roth

Main reference M. H. Kalantar, F. Rosenberg, J. Doran, T. Eilam, M. D. Elder, F. Oliveira, E. C. Snible, T. Roth, “Weaver: Language and runtime for software defined environments,” IBM Journal of Research and Development, 58(2/3):10:1–10:12, 2014.

URL <http://dx.doi.org/10.1147/JRD.2014.2304865>

We present our experiences of using Ruby as the basis for Weaver, a Domain Specific Language (DSL) to describe applications to be deployed and run in a Cloud or Cloud-like environment. Weaver describes the requirements and constituents of such an application. Based on this description, Weaver automates the deployment of such applications and facilitates the cooperation between the development and operations teams.

Weaver is an Internal DSL which means it extends the Ruby language rather than being implemented as a DSL from scratch. Extending Ruby has several advantages. For instance, we do not have to implement our own parser and additionally because our language. Additionally, our DSL shall interoperate with Chef, another Ruby-based DSL. Hence, by using Ruby, users do not have to learn yet another totally different language.

By extending an existing language we have to deal with features provided by the base language interfering with features in our DSL. One such problem is error handling, as any stack traces produced by the Ruby language will intermingle with functions provided by users as part of the DSL with functions provided by the framework used to implement the DSL.

Another advantage of Ruby is its ability to sandbox code and override Ruby’s approach for looking up symbols which allows to intercept access to variables etc. This is typically achieved with the `method_missing` method which Ruby typically invokes when a name cannot be

resolved. However, care has to be taken to distinguish legitimate calls to this method from those generated by spelling mistakes and again to factor this into errors generated by the DSL.

4.8 Probabilistic Model Checking of DTMC Models of User Activity Patterns

Oana M. Andrei (University of Glasgow, GB)

License © Creative Commons BY 3.0 Unported license
© Oana M. Andrei

Joint work of Andrei, Oana; Calder, Muffy; Higgs, Matthew; Girolami, Mark

Main reference O. Andrei, M. Calder, M. Higgs, M. Girolami, “Probabilistic Model Checking of DTMC Models of User Activity Patterns,” in Proc. of the 11th Int’l Conf. on Quantitative Evaluation of Systems (QEST’14), LNCS, Vol. 8657, pp. 138–153, Springer, 2014; pre-print available as arXiv:1403.6678v1 [cs.SE].

URL http://dx.doi.org/10.1007/978-3-319-10696-0_11

URL <http://arxiv.org/abs/1403.6678v1>

Software developers cannot always anticipate how users will actually use their software as it may vary from user to user, and even from use to use for an individual user. In order to address questions raised by system developers and evaluators about software usage, we define new probabilistic models that characterise user behaviour, based on activity patterns inferred from actual logged user traces. We encode these new models in a probabilistic model checker and use probabilistic temporal logics to gain insight into software usage. We motivate and illustrate our approach by application to the logged user traces of an iOS app. Next we will consider how to represent the orthogonal concerns of two classes of features – activity patterns and structural variability (e.g. configurability) of software systems, and their combined impact on user experience and user engagement.

4.9 Presence-Condition Simplification

Alexander von Rhein (University of Passau, DE)

License © Creative Commons BY 3.0 Unported license
© Alexander von Rhein

Joint work of Apel, Sven; Berger, Thorsten; Beyer, Dirk; Grebhahn, Alexander; Siegmund, Norbert

Analysis approaches for configurable systems take system variability explicitly into account. The notion of presence conditions is central to such approaches. A presence condition specifies a subset of system configurations in which a certain artifact is present (e.g., the presence of a certain piece of code) or any other concern of interest that is associated with this subset (e.g., the presence of a defect). Our informal goal is to raise awareness of the problem of presence-condition simplification; we will demonstrate that presence conditions often contain redundant information, which can be safely removed in the interest of simplicity. As contributions, we present a formalization of the problem of presence-condition simplification, discuss various application scenarios, compare different algorithms for solving the problem, and report on an empirical evaluation comparing the algorithms by means of a set of substantial case studies.

4.10 On the Relation between Feature Dependencies and Change Propagation

Bruno Cafeo (PUC – Rio de Janeiro, BR)

License  Creative Commons BY 3.0 Unported license
© Bruno Cafeo

As the SPL evolves, dealing with feature dependencies in the source code in a cost- and effort-effective way is challenging. It is expected that changes affect a minimum of existing features as possible. However, changes in one feature usually require changes in the code of other dependent features. In this context, it is important to have an understanding on the relation between feature dependency and change propagation. To this end, we present an exploratory study analysing this relation in five evolving SPLs. The results revealed that the extent of change propagation in SPL features might be higher than it was found in studies of change propagation in modules of stand-alone programs (i. e., non-SPL). We also found a high concentration of change propagation in a few feature dependencies. This result shows that feature dependencies are not alike regarding change propagation.

5 Breakout Groups: Domain-independence of Feature Interactions

5.1 Summary of Group 1

Krzysztof Czarnecki (University of Waterloo, CA)

License  Creative Commons BY 3.0 Unported license
© Krzysztof Czarnecki

What We've Done

- Revisited
 - Each plenary talk
 - Gerhard's presentation on systems theory
- Analyzed
 - The notion of feature used
 - The notion of feature interactions used
 - The handling of features and feature interactions in the lifecycle
- Recorded
 - Similarities and differences
 - Questions

Features

- Feature notion
 - Requirement or Behavior decomposition to understand a problem
 - Implementation unit to achieve reuse
- Purpose of features
 - Incremental development (additive) vs. independent design (“care about interactions later”)
 - Single system vs. product line (variability)
- Other characteristics
 - Automatic (prepared) composition vs. manual integration (combination)
 - Open vs. close: Are features independently developed
 - Functional vs. non-functional properties

Feature Interactions

- Vaguely: “feature behave differently together than in isolation”
 - “surprising or unexpected”, different, missing, extra behavior or properties
- Coordination code
 - Manifestation of interactions in the implementation
 - Potential exponential explosion in configurable systems
- Systems theory
 - Studying composition of components in systems and the resulting properties
 - Structure-independent properties (like mass) vs. emergent properties (like usability)

Handling Feature Interactions

- Upfront composition mechanism (architecture) vs. manual integration (combination)
 - Isolating features (e. g., Android)
- For some properties there is hope of compositionality, for some there isn’t
- Partial specifications, feature-based specifications
- Address by updates

Discussion

- Are there any conceptual problems?
- Hierarchy considered important?
- Can we learn from systems theory?
- Michael: Problem-oriented decomposition vs. component-oriented decomposition?
- Definition of feature interaction?

5.2 Summary of Group 2

Sandro Schulze (TU Braunschweig, DE), Sebastian Erdweg (TU Darmstadt, DE)

License © Creative Commons BY 3.0 Unported license
© Sandro Schulze, Sebastian Erdweg

Note: This abstract is the result of a breakout group at the Dagstuhl Seminar 14281 on Feature Interactions.

When talking about feature interactions (FIs) it becomes clear quickly that, although people work in related domains, they may have totally different views on feature interactions, which needs to be taken into account in discussions. Basically, we identified three major views on feature interactions within our breakout group. First, feature interactions may be documented in a specification such as through requirements or contracts (for example, the latter is used for verification). In any case, this provides a rather formal (and sometimes theoretical) way of defining expected FIs. Second, the source code itself may manifest a variety of feature interactions such as method calls or method extensions between two or more features. Finally, we argue that even the user may have expectations about how certain features interact, even if this expectation is not spelled out explicitly. Just think about modern cars and their capabilities to support the driver in driving the car. For instance, having a cruise control and a speed limit assistant, a driver has certain expectation how both work together, that is, how they interact.

Detecting and validating existing feature interactions is difficult without a specification that describes the expected behavior. However, during our discussion we came to the conclusion that the implementation of a feature can induce a specification beyond a formal

system or requirements specification. For instance, the requirement that a certain program (or product of a product line) does compile can be considered as a (somewhat implicit) specification. Hence, features that prevent a program from compiling in fact do constitute unwanted interactions. Other examples for such specifications are generic properties (such as absence of deadlocks or conflicts), system invariants, a poor user experience (i. e., the experienced interactions are criticized), or concrete feature specifications such as test cases. All of these specifications may support developers in detecting and validating wanted and unwanted feature interactions.

Even when we can detect feature interactions, it is even more challenging to guarantee a certain behavior of features in concert. We especially identified the fact that many (eco) systems are open world and require specific platform support for dealing with feature interactions. For example, one can deal with FIs by coordinating features on the architectural level. In other eco systems, such as the Android platform, it is common to adhere to certain conventions and to involve the user in resolving feature interactions. Finally, a practical solution to deal with feature interactions is to a) provide a default behavior in case of alternatives and b) to let the user decide which behavior she wants in case of alternatives (e. g., think of the case of selecting an App for displaying a PDF file). In any case, there is no silver bullet for how to deal with feature interactions, especially in an open-world scenario, but it is necessary to choose at least one way to resolve possible conflicts. Sometimes, this may even be a manual and time-consuming task such as informing developers about unintended FIs and (optionally) providing a corresponding patch to fix it.

To summarize, we think that feature interactions can not be treated generically, because different views and other non-functional factors have to be considered. Nevertheless, it is clear that a) there always has to be some kind of specification and b) at least an idea of how to deal with interactions when they arise. Of course, it depends on the criticality of such interactions how and when to resolve them.

5.3 Summary of Group 3

Oscar M. Nierstrasz (Universität Bern, CH)

License  Creative Commons BY 3.0 Unported license
© Oscar M. Nierstrasz

This breakout group discussed open issues and challenges for feature interaction.

We identified two essentially different perspectives on feature interaction:

1. **Design:** Here the focus is on understanding requirements with a view towards building a system in which features either do not interact, or do so in a positive way.
2. **Development:** The focus here is on code, with a view towards analysing and understanding an existing system. *Tools* are used to detect FI bugs or control quality.

We discussed three questions related to a taxonomy of FI:

What's a feature? We reviewed the results of the feature interaction survey conducted by the organizers and concluded that the two perspectives (i. e., Design vs Development) dominated: Features are either considered to be requirements (design) or units of functionality, behaviour etc. (development).

What's feature interaction? Generally, FI means that *features behave differently in isolation than in combination*. FI can take different forms: (i) features are independent and compose additively; (ii) features may depend on or require other features; (iii) combinations of

features may exhibit “emergent behaviour”; (iv) features may conflict, yielding logical inconsistencies or unacceptable behaviour.

How to deal with FIs? We identified three general approaches. (i) Modeling FI: specify desired properties (e. g., safety/liveness, non- functional properties etc.). (ii) Detecting FIs: use approaches like testing, model checking or principle component analysis to expose unknown FIs. (iii) Resolve FIs: use static or dynamic techniques, such as coordination patterns with priorities to resolve FIs.

Finally we discussed a number of open challenges and tasks for the FI community.

- Bridging the gap between the Design and Development perspectives of FI.
- Producing a map/taxonomy of FI systems and views.
- Eliciting FI patterns.
- Designing a “feature-aware computational model”, i. e., that expresses when features interact or interfere

6 Breakout Groups: Framework for Modeling Feature Interactions

6.1 Summary of Group 1

Michael Jackson (The Open University – Milton Keynes, GB)

License  Creative Commons BY 3.0 Unported license
© Michael Jackson

In this breakout session we found no reason to disagree specifically with the reference model proposed by Pamela Zave for discussion and criticism. However, one member of the group was sceptical about its possible value, on the grounds that wherever there is interaction it must have some locus: calling that locus a ‘problem domain’ added little or nothing to our understanding.

We agreed that feature interactions revealed in apparent anomalies of performance were likely to be located in a problem domain from which the analysis in hand had abstracted. So for example, the interaction might be due to conflicting demands for positioning the arm of a disk drive, where the disk drive itself was not explicitly mentioned in the immediate analysis.

It was suggested that the reference model was primarily intended as an aid to development, structuring the problem world and hence contributing to structuring the problem itself. Features could in particular cases be regarded as delimited by development or requirement modularity, or by software modules.

The correspondence between ‘subproblem machines’ and ‘behaviours’ (or perhaps ‘features’) could not be expected to carry across into software structure, for which structural transformation would surely be necessary.

Feature interaction detection by analysis of program texts seemed to some participants no different from any other formal analysis of program texts; a ‘feature’ construct in programming languages seemed highly desirable.

6.2 Summary of Group 2

Kathi Fisler (Worcester Polytechnic Institute, US)

License  Creative Commons BY 3.0 Unported license
© Kathi Fisler

One of the breakout groups discussed the idea that every feature interaction in every domain could be described as a conflict within a design domain pertinent to the system. We noted early that the hypothesis was almost certainly true, but the real question was whether this approach made practical sense: are there domains or kinds of interactions for which the design domain that witnesses the interaction is simply too fine-grained to be modeled or analyzed?

Each person in our breakout group contributed a feature interaction of interest (from a domain that that participant studies). We then selected two to discuss in more detail.

The contributed interactions included: different markup features in a text editor; synchronization policies that break upon inheritance in OO code; performance anomalies in course-management software; semantic features of an IDE that must change after the supported languages change; bugs arising from the introduction of interaction code to mediate between features; performance impacts when combining compression and encryption; call forwarding and voicemail; spam picking up email from colleagues in the contact list; and displaying filtered papers despite conflict-of-interest settings in a conference manager.

We briefly discussed the telephony example, questioning what resource was in question between call forwarding and voicemail. These simply seemed like different user-level preferences. Later discussion clarified that the single voice channel carrying the data between these two options constituted the resource in contention.

We spent much of our time debating the "change in IDE" example: we generally converged on the opinion that this was not really a feature interaction. One participant raised the idea that something counts as an interaction if some component of the system could be "blamed" for causing the problem (thus distinguishing cases of misuse from situations where interactions simply emerge through no error on the part of existing components). The group disbanded before getting to explore the blame idea in sufficient detail.

7 Panel Discussions

7.1 Reflections and Perspectives

Sven Apel (University of Passau, DE)

License  Creative Commons BY 3.0 Unported license
© Sven Apel

The panel discussion on the final day wrapped up the seminar. The format was a concentric-circles format. In the inner circle, five researchers discussed the results of the seminar, their insights, and avenues of further research on feature interactions. The panelists were: Marsha Chechik, Krzysztof Czarnecki, Michael Jackson, Christian Kästner, Pamela Zave. Joanne Atlee moderated the panel.

The panel session summarized that, during the seminar, we have seen that, in many different domains, the feature-interaction problem exists: In the telecommunication domain, the problem of feature interactions has been addressed for years, but now it appears in more

and more domains, which is owed to the fact of ever increasing complexity. An important question is whether research results can be transferred from one domain to another, and to what extent results can be domain dependent. To answer this question, the panelists suggested establishing a catalog of exemplary, domain-specific feature interactions. Examples often need to be specific to a particular domains because the domains and solutions are very different with respect to abstraction level, methodology, and requirements. The catalog is supposed to reveal – if possible – more general feature-interaction patterns, arising from concrete instances. A follow-up seminar or workshop on modeling and feature interactions shall be established to collect the data for such a catalog.

An important insight mentioned during the panelists' discussion was that we require a clearer definition of what a feature is (or should be) and how we handle features at an architectural level. To achieve better architectures, the community should investigate how to resolve and combine features independently of the domain. A catalog of feature interactions that occur in practice would be useful for this investigation.

According to the panelists, the seminar showed that there are many different aspects of feature interactions, such as the effect (good or bad), the developers' or designers' intention (intended or unintended), and the context of the interaction (design or implementation level). Several participants noted that this observation broadened their understanding of the relationship between behaviors of features in the real world and how program features work. During the discussion, it became clear that feature-oriented systems (1) are often engineered by people who are not aware of the feature-interaction problem, and (2) that are used in safety-critical domains. Therefore (intended) feature interactions must be very well explained, and we should not strive for solutions (for unintended interactions) that work only half the time. That said, features have the great potential to help people understanding complex systems in terms of the features, incl. their interactions, they provide.

Finally, the panelists stated that, while several domains experience problems similar to feature interactions, they do not call them feature interactions. An example are cars where single features are updated in a garage or even “over the air”, which give rise to feature interactions. These domains are a rich field for feature-interaction research. How can we put “the feature-interaction stamp” on them? One participant suggested having a keynote on feature interactions at a major software-engineering conference.

Participants

- Oana M. Andrei
University of Glasgow, GB
- Sven Apel
University of Passau, DE
- Joanne M. Atlee
University of Waterloo, CA
- Luciano Baresi
Politecnico di Milano Univ., IT
- Sandy Beidu
University of Waterloo, CA
- Bruno Cafeo
PUC – Rio de Janeiro, BR
- Marsha Chechik
University of Toronto, CA
- Gerhard Chroust
Universität Linz, AT
- Krzysztof Czarnecki
University of Waterloo, CA
- Nicolas Dintzner
TU Delft, NL
- Sebastian Erdweg
TU Darmstadt, DE
- Kathi Fisler
Worcester Polytechnic Inst., US
- Stefania Gnesi
CNR – Pisa, IT
- Thomas Gschwind
IBM Research GmbH –
Zürich, CH
- Reiner Hähnle
TU Darmstadt, DE
- Michael Jackson
The Open University – Milton
Keynes, GB
- Cliff B. Jones
Newcastle University, GB
- Christian Kästner
Carnegie Mellon University, US
- Mario Kolberg
University of Stirling, GB
- Sergiy Kolesnikov
University of Passau, DE
- Shriram Krishnamurthi
Brown University, US
- Malte Lochau
TU Darmstadt, DE
- Oscar M. Nierstrasz
Universität Bern, CH
- Gilles Perrouin
University of Namur, BE
- Christian Prehofer
fortiss GmbH – München, DE
- Gunter Saake
Universität Magdeburg, DE
- Sandro Schulze
TU Braunschweig, DE
- Norbert Siegmund
University of Passau, DE
- Stefan Sobernig
Universität Wien, AT
- Mirco Tribastone
University of Southampton, GB
- Alexander von Rhein
University of Passau, DE
- Andrzej Wasowski
IT Univ. of Copenhagen, DK
- Pamela Zave
AT&T Labs Research –
Bedminster, US



Crowdsourcing and the Semantic Web

Edited by

Abraham Bernstein¹, Jan Marco Leimeister², Natasha Noy³,
Cristina Sarasua⁴, and Elena Simperl⁵

1 Universität Zürich, CH, bernstein@ifi.uzh.ch

2 Universität Kassel, DE & Universität St. Gallen, CH,
leimeister@uni-kassel.de

3 Google Inc. – Mountain View, US, natashafn@acm.org

4 Universität Koblenz-Landau, DE, csarasua@uni-koblenz.de

5 University of Southampton, GB, E.Simperl@soton.ac.uk

Abstract

Semantic technologies provide flexible and scalable solutions to master and make sense of an increasingly vast and complex data landscape. However, while this potential has been acknowledged for various application scenarios and domains, and a number of success stories exist, it is equally clear that the development and deployment of semantic technologies will always remain reliant of human input and intervention. This is due to the very nature of some of the tasks associated with the semantic data management life cycle, which are famous for their knowledge-intensive and/or context-specific character; examples range from conceptual modeling in almost any flavor, to labeling resources (in different languages), describing their content in terms of ontological terms, or recognizing similar concepts and entities. For this reason, the Semantic Web community has always looked into applying the latest theories, methods and tools from CSCW (Computer Supported Cooperative Work), participatory design, Web 2.0, social computing, and, more recently crowdsourcing to find ways to engage with users and encourage their involvement in the execution of technical tasks. Existing approaches include the usage of wikis as semantic content authoring environments, leveraging folksonomies to create formal ontologies, but also human computation approaches such as games with a purpose or micro-tasks.

This document provides a summary of the *Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web*, which in July 2014 brought together researchers of the emerging scientific community at the intersection of crowdsourcing and Semantic Web technologies. We collect the position statements written by the participants of seminar, which played a central role in the discussions about the evolution of our research field.

Seminar July 6–9, 2014 – <http://www.dagstuhl.de/14282>

1998 ACM Subject Classification I.2.9 Robotics Artificial Intelligence / Robotics, D.3.1 Formal Definitions and Theory – Semantics, H.1.2 User/Machine Systems – Human information processing

Keywords and phrases Crowdsourcing, Human Computation, Games with a Purpose, Microtask Crowdsourcing, Semantic Web, Linked Data, Quality Assurance, Crowd Management, Workflow Management, Interfaces, Gamification, Incentives

Digital Object Identifier 10.4230/DagRep.4.7.25

Edited in cooperation with Cristina Sarasua



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

Crowdsourcing and the Semantic Web, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 25–51

Editors: Abraham Bernstein, Jan Marco Leimeister, Natasha Noy, Cristina Sarasua, and Elena Simperl



DAGSTUHL
REPORTS

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

1 Executive Summary

Abraham Bernstein

Jan Marco Leimeister

Natasha Noy

Cristina Sarasua

Elena Simperl

License © Creative Commons BY 3.0 Unported license
© Abraham Bernstein, Jan Marco Leimeister, Natasha Noy, Cristina Sarasua, and
Elena Simperl

The aim of the *Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web*, which was held in July 2014, was to gain a better understanding of the dual relationship between crowdsourcing and Semantic Web technologies, map out an emerging research space, and identify the fundamental research challenges that will need to be addressed to ensure the future development of the field.

The seminar focused on three categories of topics: first and foremost we looked into existing crowdsourcing approaches and how these could or have been applied to solve traditional semantic data management tasks. Particular attention was paid to core components of a crowdsourcing-enabled data management and processing system, including methods for quality assurance and spam detection, resources, task and workflow management, as well as interfaces, and the way these components can be assembled into coherent frameworks. A second category of topics that was addressed during the seminar reached out to other disciplines such as economics, social sciences, and design, with the aim to understand how theories and techniques from these fields could be used to build better crowdsourcing-enabled data management systems for the Semantic Web. Last, but not least, we discussed the usage of semantic technologies within generic crowdsourcing scenarios, most notably as means to describe data, resources and specific components.

The seminar, in its community-formative role, represented the starting point for the emergence of working groups that will in the future jointly address the identified scientific challenges. Participants were asked to provide a 1-page position statement reflecting on why they think it makes sense to consider the two topics – crowdsourcing and Semantic Web (or Web of Data) – at the same seminar. Specifically, participants were asked to write a statement reflecting on one of both of the following questions:

1. What are the Semantic Web tasks where you felt you needed crowdsourcing? Why? What were the challenges?
2. What are the crowdsourcing tasks where using semantics might help? Why? What are the challenges?

The first two days of the seminar were dedicated to presentations of topics related to position statements and working groups on use case scenarios and challenges identified during the talks and Q&A sessions. The third day focused on the consolidation of the results of the working groups and the definition of next steps and follow-up activities.

In the following sections we present the position papers written by the researchers of the crowdsourcing and the Semantic Web community, who took part in the seminar. We will publish a more complete research roadmap for crowdsourcing and the Semantic Web at a later stage.

2 Table of Contents

Executive Summary

<i>Abraham Bernstein, Jan Marco Leimeister, Natasha Noy, Cristina Sarasua, and Elena Simperl</i>	26
--	----

Position Papers

Crowdsourcing Linked Data Management <i>Maribel Acosta</i>	29
Using Crowdsourcing for Semantic Annotation in Media Sector <i>Sofia Angeletou</i>	30
Semantic Interpretation and Crowd Truth <i>Lora Aroyo</i>	31
Check and Track – Crowdsourcing Semantics and Semantic Crowdsourcing <i>Irene Celino</i>	31
Crowdsourcing & the Semantic Web <i>Philippe Cudré-Mauroux</i>	32
Crowdwork and Microtasks <i>Roberta Cuel</i>	33
Crowdsourcing is for the Tail <i>Gianluca Demartini</i>	34
Crowdsourced Feature Selection <i>Michael Feldman</i>	35
Crowdsourcing Semantic Tasks for Scientific Research <i>Yolanda Gil</i>	36
Position Statement Dagstuhl Seminar on Crowdsourcing and the Semantic Web <i>Carole Goble and Steve Pettifer</i>	37
Crowdsourcing Platforms and the Semantic Web <i>Atsuyuki Morishima</i>	38
Crowdsourcing and the Semantic Web one-page Position Statement <i>Valentina Presutti</i>	39
Crowdsourcing Ontology Lexicalization <i>Philipp Cimiano</i>	40
Towards Hybrid-genre and Embedded Crowdsourcing <i>Marta Sabou</i>	41
Crowdsourcing for Evaluation and Semantic Annotation <i>Harald Sack</i>	43
Who and How Should Be Involved in Crowdsourced Data Interlinking? <i>Cristina Sarasua</i>	44
Linking Implicit with Explicit Semantics: An Initial Position Statement <i>Markus Strohmaier</i>	46
Opinions and Aims in Participatory Sensing <i>Gerd Stumme</i>	46

Crowdsourcing and the Semantic Web	
<i>Tania Tudorache</i>	47
The Role of Crowdsourcing and Semantic Web for Consumable APIs	
<i>Maja Vukovic</i>	48
Training Systems with Crowd Truth	
<i>Christopher A. Welty</i>	49
Merging Contexts	
<i>Marco Zamarian</i>	50
Participants	51

3 Position Papers

This section includes the complete list of position statements, which participants (except for the organizers) provided.

3.1 Crowdsourcing Linked Data Management

Maribel Acosta (KIT – Karlsruher Institut für Technologie, DE)

License  Creative Commons BY 3.0 Unported license
© Maribel Acosta

The Semantic Web is envisioned as a system in which machines can truly understand the meaning of requests posed by humans or other machines, assuming that all the data consumed and produced in this system is encoded with semantics. For this vision to become a reality, it is necessary in the first place to create semantically enriched data that will allow machines to have such comprehension of the data. Furthermore, even in the presence of semantic data, there are still data enhancement tasks to be addressed that require the execution of processes that are intrinsically better performed by humans than by machines, like disambiguation, association, pattern recognition, etc. In particular, considering the limitations of machines when the meaning of the data is highly contextual or subjective, we can foresee a powerful use of crowdsourcing approaches for Linked Data management tasks for the following problems:

Link prediction

Involves the creation of links on the instance level. Link prediction includes the problem of entity resolution by creating owl:sameAs links as well as the creation of other type of links between resources. This is particularly challenging when the data lacks specificity (no useful additional information is provided), has many homonyms (requiring a disambiguation step), and its variety is high. Therefore, the decision to create a link or not is highly influenced on the context and type of the data.

Data quality assessment

Refers to the process of validating data to detect inconsistencies or other type of errors in the data. In order to successfully execute this task, it is necessary to discern the possible types of errors encountered in the data to perform the appropriate corrective actions.

In the previously discussed tasks, human intervention can serve different purposes. For instance, crowdsourcing approaches could perform all the single steps of a task. The main challenge in this case is to ensure scalability in terms of monetary cost and execution time when executing very large tasks. Another strategy would be to apply crowdsourcing to generate training data, e.g., ground truth datasets, and implement supervised learning techniques. While this approach is more flexible in terms of scalability, since learning approaches are highly sensitive to training data, this strategy requires the creation of high quality data from the crowd. As a complement, crowdsourcing techniques can also be applied to validate intermediary outcomes of automated approaches.

3.2 Using Crowdsourcing for Semantic Annotation in Media Sector

Sofia Angeletou (BBC – London, GB)

License  Creative Commons BY 3.0 Unported license
© Sofia Angeletou

One of the challenges in the uptake of semantic technologies and linked data is the lack of high quality, semantically annotated content that can be reused and repurposed using linked data principles. Although there is a plethora of open vocabularies available, the volume of good quality annotated content across various domains is not comparable. Such a gap hinders the creation of applications that could make use of such content and showcase the value of this technology in real world use cases; either in the public domain (open data) or internally in organisations.

In this paper I argue that using crowdsourcing is a means to obtaining high quality semantically annotated content with a special focus on the Media sector. The primary output of organisations in this sector includes the publication of content assets such as programmes, news articles and educational works both in a linear broadcasting but also on an “on demand” basis. The cases for using semantically annotated content vary from improved management of media assets, creation of curated content indexes with a strong semantic layer and novel audience facing applications that involve personalised content offerings based on the things that matter to each individual member of the audience.

Current observations in some Media organisations show that semantic annotations are inconsistent both in completeness and in quality, rendering the quality of the consuming product poor. The semantic tagging techniques may involve entity extraction as a first step, but then rely on a single editor to apply or approve relevant tags. This does not always yield a good tagging result, given the subjectivity of the tagger and various workflow factors, such as availability of time available for annotation. Having the manual annotations weighted based on the popularity of selected concepts can contribute to the solution of the problem to a large extent. Crowdsourcing could be encouraged either internally in an organisation or opened up to the public, to allow interested volunteers contribute to the annotations. However, both cases pose interesting challenges.

In internal crowdsourcing, the process should be designed such that it fits in the workflow of content editors without creating additional overhead, given their restricted time. In addition they should be informed of the impact of their action as a motivation to continue contributing. In many cases the value of tagging is not clear to the content editors or annotators, it is often seen as an additional and disconnected task they must complete that causes delays in the accomplishment of other production activities.

In cases where crowdsourcing could be employed in an open world setting there are a few challenges that would influence both the result but also the perception of organisations about implementing such practices. Some forms of crowdsourcing would typically involve technically competent users of a service, who might be biased towards particular selections not representative of the whole of the audience of an organisation. In addition, allowing the public to annotate content would require further editorial control to ensure that the annotation results reflect the editorial policies and branding of the organisation.

3.3 Semantic Interpretation and Crowd Truth

Lora Aroyo (*Free University of Amsterdam, NL*)

License  Creative Commons BY 3.0 Unported license
© Lora Aroyo

Human annotation is a critical part of semantic processing, from the early days of knowledge acquisition to modern methods of collecting data to train and evaluate machine learning algorithms that perform semantic interpretation tasks. However, conventional human annotation, and indeed Semantic Web technologies themselves, are based on an antiquated ideal of a single correct truth that needs to be disrupted. I propose a new theory of truth, Crowd Truth, that is based on the intuition that human interpretation is subjective, and that measuring annotations on the same objects of interpretation (e.g. sentences, images, videos) across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations. I will introduce several metrics for measuring quality of human annotated data based on Crowd Truth, and present experimental results that show these metrics are inter-related in a way that previous human annotation methodologies did not reveal.

3.4 Check and Track – Crowdsourcing Semantics and Semantic Crowdsourcing

Irene Celino (*CEFRIEL – Milano, IT*)

License  Creative Commons BY 3.0 Unported license
© Irene Celino

The Semantic Web needs Crowdsourcing for a number of different tasks; however, in my opinion, the most fruitful role for Crowdsourcing is for data quality and validation in the Semantic Web. The vast popularity and success of the Linked Data initiative has been bringing an increasingly large amount of data on the Semantic Web or Web of Data, often in form of raw data or automatically-generated/transformed information. Thus, the issue of checking data quality, correctness, consistency and update becomes of utmost importance [7, 6, 3] and here is where I see a concrete use of Crowdsourcing techniques. In my experience of building Human Computation games for geo-spatial data management [2], the results showed that data curation (especially in case of outdated information) was definitely a plus with respect to pure data collection, in terms of precision and outcome value. In the context of Semantic Web research, Crowdsourcing has been employed also for other knowledge-intensive tasks (e.g. ontology engineering or ontology alignment [5]), at times also successfully. Still, I believe that the simplicity of the crowdsourced task should be always taken into due consideration, thus data curation could be a better purpose for Crowdsourcing rather than ontology management; in any case, the issue of selecting the right crowd is always open. Moreover, fact checking is definitely a domain in which human judgement easily outperforms automatic techniques, so it is probably a more relevant objective for applying Crowdsourcing techniques.

Crowdsourcing needs the Semantic Web for crowd users' tracking: the issue of understanding who the crowd workers are, their background and expertise, and ultimately their reliability is central to any Crowdsourcing effort. This is why I believe that Semantic Web technologies can be key to describe people profiles as well as to track the provenance of an information

value chain (i. e., workflow of the data lifecycle). Indeed, models derived or inspired by W3C PROV-O have been adopted to keep trace of human interventions [1] or to log modifications to a triple dataset [4] (collected both via crowdsourcing and heuristic/statistical approaches). As a consequence, Semantic Web techniques can be employed to supervise the Crowdsourcing process, in terms of evidence collection, agreement and decision making, and especially crowd users' tracking and evaluation. Tracing the crowdsourced information can be very useful to compare different strategies to aggregate results as well as to provide incentives and rewards to the crowd.

References

- 1 I. Celino. Human Computation VGI Provenance: Semantic Web-based Representation and Publishing. *IEEE Transactions on Geoscience and Remote Sensing*, 51(11):5137–5144, 2013.
- 2 I. Celino. Geospatial Dataset Curation through a Location-based Game. *Semantic Web Journal*, 5(7), 2014.
- 3 I. Celino, E. Della Valle, and R. Gualandris. On the effectiveness of a Mobile Puzzle Game UI to Crowdsourcing Linked Data Management tasks. In *1st International Workshop on User Interfaces for Crowdsourcing and Human Computation*, 2014.
- 4 M. Knuth and H. Sack. Data Cleansing Consolidation with PatchR. In *Posters and Demos – ESWC 2014*, 2014.
- 5 C. Sarasua, E. Simperl, and N. Noy. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *11th International Semantic Web Conference*, pages 525–541, 2012.
- 6 E. Simperl, B. Norton, and D. Vrandečić. Crowdsourcing Tasks within Linked Data Management. In *Proceedings of COLID2011*, volume 782. CEUR-WS.org, 2011.
- 7 J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. Whoknows? evaluating linked data heuristics with a quiz that cleans up dbpedia. *Interactive Technology and Smart Education*, 8(4):236–248, 2011.

3.5 Crowdsourcing & the Semantic Web

Philippe Cudré-Mauroux (University of Fribourg, CH)

License  Creative Commons BY 3.0 Unported license
© Philippe Cudré-Mauroux

The Semantic Web was, from its inception, destined to create a machine-processable Web of data, a Web where computerized agents could collect, integrate, exchange and reason upon large quantities of heterogeneous online information. Over the years, however, new applications of the Semantic Web emerged. Today, some of its most exciting applications—such as the display of semi-structured information related to an entity, or the parsing of natural language for text summarization or question answering—are directly targeting human users. In that sense, the Web of data is increasingly used to help humans in their daily lives. In contrast, crowdsourcing leverages a Web of documents, made for humans, to help machines solve computationally complex tasks. Crowdsourcing and the Semantic Web were in that sense bound to meet each other, and to come together to solve some of the most formidable open problems in information management. In my research group, the eXascale Infolab¹ both topics frequently support each other. On one hand, we recently used crowdsourcing to help solve complex Semantic Web issues such as entity linking [1] or instance matching [2]. Semantic

¹ <http://exascale.info/>

data, on the other hand, was instrumental in facilitating advances in push crowdsourcing [3] and in crowdsourced data sensing [4]. Another interesting convergence might occur in a few years, when generalized micro-task crowdsourcing platforms will emerge and host arbitrary complex tasks—some annotated using Semantic Web information. Both human and machines might then compete for the same tasks on the crowdsourcing infrastructure, creating de facto and for the first time a universal and hybrid Semantic Web service infrastructure for information processing.

References

- 1 G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In WWW, pages 469–478, 2012.
- 2 G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. VLDB J., 22(5):665–687, 2013.
- 3 D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In WWW, pages 367–374, 2013.
- 4 M. Wisniewski, G. Demartini, A. Malatras, and P. Cudré-Mauroux. Noizcrowd: A crowd-based data gathering and management system for noise level data. In MobiWIS, pages 172–186, 2013.

3.6 Crowdwork and Microtasks

Roberta Cuel (University of Trento, IT)

License  Creative Commons BY 3.0 Unported license
© Roberta Cuel

Organizations increasingly use crowdsourcing to solve problems outside the reach of traditional work processes. By having access to a practically unlimited pool of contributors providing an unprecedented spectrum of skills, experiences and ideas, organizations are enabled to organize their knowledge, increase the flexibility of their processes, and drive openinnovation.

Crowdsourcing for Semantic Web

Organizations can use crowdsourcing to organize their unstructured information and knowledge, identifying entities and disambiguating meaning of words, images, videos, etc. This semantic knowledge can later be used to improve searching and reasoning processes in the corporate knowledge memories, knowledge bases, archives, etc. This development comes with a multitude of social, legal, technical and economic challenges. In this page I want to focus on two of these. How people understand and find an agreement upon a piece of information structure. Especially in big corporates, workers belonging to different units (R&D, production, marketing, accounting, etc.) use different interpretation schemas to identify entities and disambiguate meaning. Indeed, depending on different interpretation schemas, people may use the same categories with different meanings or different words to mean the same thing. An entity can be considered as the “explicit part of what we know” and gets its meaning from a typically implicit taken for granted interpretation schemas (among others see paradigms [Kuhn, 1979], frames [Goffman, 1974], thought worlds [Dougherty, 1992], context [Ghidini & Giunchiglia, 2001], mental spaces [Fauconnier, 1985], cognitive path [Weick, 1979], etc.) How workers deal with crowdsourcing and their daily activities. People’s participation and willingness to contribute are critical issues that organizations

must take into account when introducing crowdsourcing solutions. In the specific case of a corporate setting, we are in a principal agent relationship in which the two parties (employer and workers) have different interests and asymmetric information. In order to ensure high quality results and reduce moral hazard and conflict of interests a set of incentive should be designed.

Semantic Web for Crowdsourcing

Semantic web can enable crowdsourcers (content providers) to provide a better quality contributions. In the particular case of microtasks, such as document/image annotations and very simple semantic disambiguation, crowdsourcers can take advantage from the suggestions provided by a semantic based system. We developed a semantic based platform for the disambiguation of Diagnosis-related groups (DRG) in an hospital setting. Doctors have improved their contributions, getting suggestions about what DRGs better suite the description of a patient record. In that case, Semantic Web helped doctors to choose a concept among a pre-elaborated set of entities, reducing time of patient record elaboration and DRG recognition, improving the precision of the choices and the variability in the DRG recognition. One of the big challenges is the process of knowledge convergences or divergences that may emerge do to the suggetions provided by the semantic based system. When people have to choose among a selection of pre-defined items, tend to use them as they are, and the final result is often a common and unique conceptualization of knowledge, not very innovative.

References

- 1 E. Von Hippel, Horizontal Innovation Networks U by and for Users, working paper 4366-02, MIT Sloan School of Management, 2002.
- 2 R. Cuel et al., Motivation Mechanisms for Participation in Human-Driven Semantic Content Creation, Int'l J. Knowledge Eng. and Data Mining, vol. 1, no. 4, 2011, pp. 331–349.
- 3 N. Kaufmann, T. Schulze, and D. Veit, More than Fun and Money, Worker Motivation in Crowdsourcing U A Study on Mechanical Turk, Proc. 17th Americas Conf. Information Systems, AIS, vol. 1, no. 11, 2011, pp. 1–11.
- 4 C. Prendergast, The Provision of Incentives in Firms, J. Economic Literature, vol. 37, no. 1, 1999, pp. 7–63.
- 5 D. B. Brabham, Moving the Crowd at Threadless: Motivations for Participation in a Crowdsourcing Application, Information, Communication, and Society, vol. 13, no. 8, 2010, pp. 1122–1145.

3.7 Crowdsourcing is for the Tail

Gianluca Demartini (University of Fribourg, CH)

License  Creative Commons BY 3.0 Unported license
© Gianluca Demartini

Semantic Web needs crowdsourcing for data quality issues. However, the seminar should just briefly overview the standard and obvious tasks like: Entity linking (attach URIs to entities in text), knowledge base integration (identify the same entity over two KBs and create owl:sameAs statements), ontology matching (identify the same predicates in two ontologies), relation creation/check (generating or validating RDF predicates), and knowledge base quality curation (fact checking). For sure, within this set of tasks an open question is how

to optimally design HITs for the crowd trying, for example, to avoid Semantic Web specific terminology like ontology, predicate, URI, etc.

The aspect I would like to discuss the most is to, which, data crowdsourcing should be applied in order to make hybrid human-machine approaches suitable at Web scale. Human computation approaches (either crowdsourcing or editorial curation) are already successfully applied to very popular entities: Examples include the Google Knowledge graph and Wikipedia where either employees or crowds manage the content and improve its quality. The open challenge remains for tail, entities, that is, the very many different entities which are not popular or valuable enough individually but would have a great value as a whole (e. g., collecting all small restaurants opening hours in a city).

The point I want to make at the seminar is that micro-task Crowdsourcing has to be used for tail entities and not just for entities worth to appear in Wikipedia or in the Google Knowledge Graph. In order to make the crowd effectively work on those entities we need to leverage worker skills and passions [1] or the communities they belong to [2]. The question then becomes about which crowd should be used for a specific task. To better answer this question, that is, to find the right workers in the crowd for a task, we need to leverage profiling techniques, recommender, systems, and push, crowdsourcing as we are currently doing with OpenTurk [3]. In other cases when the right worker is not available, it is necessary to train the crowd before it can perform some tasks, as, for example, dealing with domain specific (e. g., medical) entities.

References

- 1 Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do. In: 22nd International Conference on World Wide Web (WWW 2013), Rio de Janeiro, Brazil, May 2013.
- 2 Michele Catasta, Alberto Tonon, Djellel Eddine Difallah, Gianluca Demartini, Karl Aberer, and Philippe Cudré-Mauroux. Hippocampus: Answering Memory Queries using Transactive Search. In: 23rd International Conference on World Wide Web (WWW 2014), Web Science Track. Seoul, South Korea, April 2014.
- 3 <http://alpha.openturk.com>

3.8 Crowdsourced Feature Selection

Michael Feldman (Universität Zürich, CH)

License  Creative Commons BY 3.0 Unported license
© Michael Feldman

I am currently interested in feature selection for classification tasks based on wisdom of crowd. So far this domain was a prerogative of machine learning algorithms, responsible for extracting most substantial features such that by knowing them the entity will be classified with minimal error [3]. However, while this approach is effective for explicit and structured data, extracting features to classify non-trivial, logical structures is extremely challenging for traditional machine learning algorithms [2], [1]. For instance, to classify writer by her writing style may be done by natural language processing methods as Latent Semantic Analysis. However, these methods have limited success with significant drawbacks as scalability, synonymy or polysemy treating. Therefore I hope to explore during the workshop the following aspects:

1. Challenging Semantic Web tasks that can not be resolved efficiently with existing tools, but the performance of such tools may be boosted by extracting tacit knowledge of crowds.

2. Exploring the existing methods of crowd engagement to solve typical Semantic Web problems.
3. Possible ways to explore the tacit knowledge of crowds and to outline it in succinct way (e. g., as a set of features or significant patterns) with regards to relevant Semantic Web tasks. All this, taking in account existing Semantic Web tools and ability to verify the extracted data contribution by boosting performance of these tools.
4. To gain an understanding of how to conceptualise Semantic Web tasks to general framework. As tasks may differ in their definitions (e. g., writing style vs. data integration), I would like to explore approaches providing the means to define general problem while parameters of the tasks are different.

References

- 1 Alelyani, S., Tang, J., Liu, H.: Feature selection for clustering: A review. *Data Clustering: Algorithms and Applications* p. 29 (2013).
- 2 Liu, H., Motoda, H., Setiono, R., Zhao, Z.: Feature selection: An ever evolving frontier in data mining. *Journal of Machine Learning Research-Proceedings Track 10*, 4–13 (2010).
- 3 Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. *Knowledge and Data Engineering, IEEE Transactions on* 25(1), 1–14 (2013).

3.9 Crowdsourcing Semantic Tasks for Scientific Research

Yolanda Gil (Information Sciences Institute, University of Southern California, US)

License  Creative Commons BY 3.0 Unported license
© Yolanda Gil

We are investigating the use of semantic wikis to develop platforms that enable the flexible incorporation of contributors into science tasks. Our research in this area is in three major fronts:

1. **Organic Data Curation:** There are many datasets in geosciences that have been collected over the years. Even if the datasets are available in centralized repositories, they rarely have metadata. Organizing and describing these datasets are great candidate tasks for crowdsourcing. We are investigating how to form communities that can create metadata for repositories that contain thousands of scientific datasets. We are developing an organic data curation framework that extends semantic wikis so that the contributors can dynamically define new metadata attributes on the fly, and to use those attributes to describe the properties of datasets. We are incorporating in our framework the incentives and rewards that work best for scientific researchers and citizen volunteers to contribute to these tasks.
2. **Organic Data Science:** According to some estimates, 85% of geosciences data is often kept by individual investigators and not available in shared repositories, often called “darked data.” Many global environmental science research cannot be carried out without accessing this data. We are developing an organic data science framework where the task of sharing and curating data is an integral part of doing science, and where data contributors can participate in tasks that involve the use of their data.
3. **Tracking the Use of Semantic Wikis in Science:** Semantic wikis have been extensively used to collect and structure scientific knowledge. We are analyzing the content and growth of these wikis as platforms for doing science using Semantic Web technologies. We

have developed ProvenanceBee, a service that collects semantic content and provenance of hundreds of semantic wiki sites. The semantic wiki platform is used very differently in different scientific applications. Our research focuses on studying semantic wiki communities to understand how contributors add semantic content to wikis.

3.10 Position Statement Dagstuhl Seminar on Crowdsourcing and the Semantic Web

Carole Goble and Steve Pettifer (University of Manchester, GB)

License  Creative Commons BY 3.0 Unported license
© Carole Goble and Steve Pettifer

What are the Semantic Web tasks where you felt you needed crowdsourcing?

Making scientific knowledge in the literature accessible to machines by stealth.

Why? Currently the knowledge is captured in databases and the scientific literature; these are growing at a rate that makes it impossible for humans to keep up with latest info without support from computers to find and summarize important discoveries.

Databases have at least some structure and are intended to be machine-readable, so there is scope for converting these into formats that are friendlier to the Semantic Web. The literature on the other hand is largely designed for humans; it is full of ambiguities, subtle nuances and rhetorical flourishes that work for human readers but confuse machines. This, and the fact that much of the literature is stored in formats that are hard to process by machine, makes extracting semantics from articles very hard.

What were the challenges

- Much important information is captured in natural language, diagrams or tables. These are hard for a computer to process.
- Legacy formats such as the PDF are typically un-semantic bags of words and lines/curves are difficult to process reliably into anything structured.
- The hedged language used in scientific writing to avoid over-claiming makes identifying claims hard (i.e. we don't say "A does B to C", but "We hypothesize that it may be the case that A, under certain circumstances might behave in a B-like way in the context of C").

What are the crowdsourcing tasks where using semantics might help?

Paywalls and licenses make bulk-mining of the literature by an individual hard/impossible/illegal, but nothing prevents a single person mining a single paper to which they have legitimate access. Doing this en-masse and combining the results would yield a valuable body of machine processable knowledge.

Why? However, text and data mining algorithms are heuristic; semantics are necessary to normalize / cross-validate results to gain confidence, and to manually correct/curate errors in the automatic extraction.

What are the challenges? It is difficult to persuade busy scientists to do work that is for the collective good without any immediate payback; therefore micro tasks have to be built

into a system where the user is either a ‘side effect’ of something they want to do anyway, or is low-effort and has obvious immediate benefit to them.

Project Lazarus

The purpose of Project Lazarus [1] is to attempt to crowd-source information from the scientific literature ‘by stealth’, as a side effect of providing scientists with a tool that makes navigating and exploring the scientific literature easier and more rewarding than by conventional means. Scientists thus gain an immediate benefit, and as a side effect of using the tools contribute to a central repository of semantically rich and openly licenced knowledge.

Utopia Documents [2] is a PDF reader that provides many immediately useful features; for example the ability to dereference and ‘click through’ references to retrieve the cited article without needing to search online, or to extract data from tables into spreadsheets. Project Lazarus aims to extend the tool to capture the ‘exhaust gasses’ of such activities. With a user’s permission, the data from these actions can automatically be contributed to a central repository creating (for example) a citation network, or a repository of data-from-tables. The system can automatically associate provenance with the data, i.e. which user did the extraction, using which algorithm, when, and from which scientific article.

References

- 1 BBSRC BB/L005298/1 The Lazarus Project: Resurrecting data and knowledge from life science articles by crowd-sourcing, <http://www.bbsrc.ac.uk/pa/grants/AwardDetails.aspx?FundingReference=BB/L005298/1>.
- 2 <http://getutopia.com>

3.11 Crowdsourcing Platforms and the Semantic Web

Atsuyuki Morishima (University of Tsukuba, JP)

License  Creative Commons BY 3.0 Unported license
© Atsuyuki Morishima

Crowdsourcing is a promising tool to solve some of the problems in the semantic Web domain. We are operating a crowdsourcing platform named Crowd4U with the help of researchers from more than 22 universities and conducting several projects for the academic and public purposes. In this paper, we first explain Crowd4U and a project running on it, named L-Crowd, in which we try to clean bibliographic data by crowdsourcing performing tasks for entity identification in bibliographic records. Then, we discuss lessons learned that might be related to the Semantic Web.

Crowd4U

Crowd4U is a microtask-based crowdsourcing platform which is similar to Amazon mechanical turk. It is unique in several ways compared to similar systems. First, Crowd4U provides a high-level abstraction for complex human-machine computation. Second, it supports various task assignment and incentive structures including push/pull-style task assignments. Because many workers voluntarily perform tasks on Crowd4U, they are called *contributors*. Many of these contributors are university students. They performed more than 1,100 tasks per day

in May 2014. The estimated number of anonymous workers on Crowd4U is more than one thousand.

L-Crowd

Crowd4U is hosting several non-commercial crowdsourcing projects.

L-Crowd is one of these projects that started by LIS and CS researchers in Japan to apply crowdsourcing technologies to library problems. In 2012, they designed microtasks to identify different books that have the same ISBN in an effort to clean the bibliography database of the National Diet Library

Lessons Learned and Challenges

From our experience of operating a crowdsourcing platform and conducting several crowdsourcing projects, we think the followings are important related to the Semantic Web.

(1) What are the Semantic Web tasks where we felt we needed crowdsourcing? Entity identification is definitely an important task that needs crowdsourcing. It requires workers to understand the context and conduct some inference based on the background knowledge. Another important task might be to connect concepts to each other in different languages.

(2) What are crowdsourcing tasks where using semantic might help? First, semantic can help workers understand the instructions in the task. Second, semantic can show workers possible results for the task so that workers can choose one of them instead of thinking up their own answer. Third, semantic can improve to data quality of the task results, because it can identify unlikely results for the tasks.

From our point of view, an important challenge is to identify fundamental functions that crowdsourcing platforms should provide to support Semantic Web applications and to use semantic to support effective crowdsourcing.

3.12 Crowdsourcing and the Semantic Web one-page Position Statement

Valentina Presutti (CNR – Rome, IT)

License © Creative Commons BY 3.0 Unported license
© Valentina Presutti

What are the Semantic Web tasks where you feel you need crowdsourcing?

- Building cognitive-based resources: resources such as WordNet, FrameNet, etc. should be built and/or validated by means of crowdsourcing in order to reflect the way humans cognitively organize and use their knowledge;
- User-based evaluations: for evaluating methods supporting cognitive-oriented tasks, e.g. content/entity summarization, exploratory search (serendipity, discovery), knowledge relevance, etc., as well as more technical tasks, e.g., formal representation of natural language, ontology learning, property and concept alignment, etc.;

Why?

- because statistical significance requires good numbers of users involved;
- because cognition is a human feature;
- because we need to understand human cognitive capability;

- because it would save time;
- because we need to cope with subjective (context-based) perspectives;
- because I think golden standards-based evaluations often do not apply to Semantic Web research.

What are the challenges?

- identifying the right crowd;
- will to commit to perform high-quality work;
- ensuring high-quality standards (e. g. extremely important for cognitivebased resources);
- reducing complex tasks to simple ones: can we conceive a sort of “reduction” method (cf. complexity theory) to formally or at least rigorously define simplified versions of SW tasks to be assigned to the crowd;
- developing methods that ease modeling crowdsourcing tasks;
- finding incentives for experts to commit to crowdsourcing;

3.13 Crowdsourcing Ontology Lexicalization

Philipp Cimiano (Bielefeld University, DE)

License  Creative Commons BY 3.0 Unported license
© Philipp Cimiano

For all those applications in which natural language is used to access and interface Semantic Web data, knowledge is needed about how the vocabulary used to describe the data is expressed in natural language. The current state of affairs is one in which each application needs to derive the relation between natural language and formal vocabulary anew. To avoid this situation, lexicon models such as the Lexicon Model for Ontologies (lemon) [3] or the Ontology Lexicon Model (ontolex) have been and are being developed, allowing one to represent this knowledge in a declarative fashion, thus supporting the sharing of this knowledge across applications. Nevertheless, such lexicon models need to be populated, which is a costly process, as different variants of how to refer to a particular vocabulary element in a particular language need to be included. We refer to the task of specifying all the different (lexical) variants that can be used to refer to a particular vocabulary element in some vocabulary or ontology as ontology lexicalization. As an example, consider the property `<http://dbpedia.org/ontology/spouse>` in DBpedia. This property can be verbalized as follows in English:

X is married to Y X married Y
 X is the spouse of Y X is the wife of Y
 X is the husband of Y
 X is the better half of Y, etc.

As a proof-of-concept of the lemon lexicon model, a manual lexicon for DBpedia has been created [1]. So far, semi-automatic approaches to induce ontology lexica from a corpus have been proposed [2]. In spite of using a corpus-based approach, the work of Walter et al. has shown that human validation is still needed to reach an appropriate quality of the final resource. This, however, is an issue, as human labour needs to be found and rewarded for their work in some way or another (either through payment or other incentives). Crowdsourcing might play a key role in this, involving workers in both the specification of lexical variants as well as the validation of entries proposed by an automatic approach such as the one by Walter et al. Relevant research questions in this context are the following ones:

1. How might a crowdsourcing-based framework for ontology lexicalization look like?
2. How should the tasks be formulated to maximize effectiveness and efficiency?
3. By which heuristics or automatic checks can overall quality be ensured?
4. What skills are needed by workers to accomplish the task?
5. What is the cost of lexical pattern acquisition for each vocabulary element?
6. How can workers be motivated to perform the task other than via (micro-) payments?
7. Would a methodology working for English scale to other languages?

References

- 1 Christina Unger, John McCrae, Sebastian Walter, Sara Winter, Philipp Cimiano: A lemon lexicon for DBpedia. In: Proc. of the NLP & DBpedia Workshop, collocated with ISWC 2013.
- 2 Sebastian Walter, Christina Unger, Philipp Cimiano: A Corpus-Based Approach for the Induction of Ontology Lexica. In: Proc. of the 18th Int. Conf. on Applications of Natural Language to Information Systems (NLDB), 2013.
- 3 John McCrae, Dennis Spohr, Philipp Cimiano: Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In: Proc. of the Extended Semantic Web Conference (ESWC) 2011.

3.14 Towards Hybrid-genre and Embedded Crowdsourcing

Marta Sabou (MODUL Universität Wien, AT)

License  Creative Commons BY 3.0 Unported license
© Marta Sabou

We see (1) cross-genre crowdsourcing and (2) a tighter integration of crowd-work into ontology engineering as exciting challenges of applying crowdsourcing in the Semantic Web area. Many open issues relate to semantically describing crowd-worker profiles, crowdsourcing tasks, data and workflows to enable advanced functionalities such as (1) flexible matchmaking between tasks and workers or (2) automatic discovery and combination of complex crowdsourcing workflows.

Crowdsourcing for Semantic Web

The Semantic Web task we aimed to solve by crowdsourcing was creating high quality domain ontologies based on an initial semantic structures extracted automatically by an ontology learning algorithm. We needed multiple crowdsourcing tasks, which can be grouped into two categories:

1. Verifying extracted content: are the extracted concepts correct and relevant for the domain? are the extracted subsumption (and other types of) relations correct?
2. Creating new content that is still challenging to extract automatically: given two concepts, is there a relation between them? and if yes, what is that relation?

To accomplish these tasks, we used both games with a purpose (GWAPs) and paid-for crowdsourcing approaches, concluding that both genres have their pros and cons [1]. GWAPs require a higher upfront investment for designing and building them, but the collected contributions are free. Challenges include: designing appealing games; attracting and maintaining a high number of players. In contrast, mechanised labour tasks take less time and effort to set up, can be outsourced to large pools of workers with highly diverse

qualifications, but each contribution costs. On the down-side, it is challenging (1) to phrase tasks in ways that are understood by layman; (2) to ensure the quality of the collected data (e. g., through task design, gold data etc); (3) to find the optimal task setups that maximizes quality while reducing overall cost and completion times. We highlight the following research challenges in applying crowdsourcing for the Semantic Web:

How Can Multiple Crowdsourcing Genres Be Combined? Given the complementary strengths and weaknesses of the GWAPs and paid for crowdsourcing mechanisms, is it possible to combine them in a beneficiary way? For example, crowdsourcing tasks could be split into workflows of tasks, where simpler (e. g., verification) tasks are crowdsourced for money while challenging (e. g., content creation) tasks are solved through game playing. We proposed hybridgenre crowdsourcing as a potential approach [2], but many open issues remain.

How to Embed Crowd-work Into Ontology Engineering? The use of crowdsourcing in the Semantic Web community has matured enough to move on from isolated approaches towards a methodology of where and how crowdsourcing can efficiently support ontology engineers. We see the derivation of such methodologies as an important prerequisite for popularizing the use of crowdsourcing by ontology engineers. Such methodological guidelines should inform the development of tools that facilitate easily embedding crowdsourcing into ontology engineering workflows (for example, extensions of ontology editors).

Semantic Web for Crowdsourcing

Matchmaking between crowd-workers and tasks is an increasing challenge as crowdsourcing platforms attract larger worker bases with diverse qualifications and crowdsourcing is used for novel and diverse tasks. Semantically representing worker profiles and task types in order to create an optimal matchmaking between them could be an interesting task to solve using Semantic Web technologies. This would allow migrating from pull- to push-based crowdsourcing mechanisms. A key challenge here lies in determining a meaningful category of tasks and finding large enough platforms where these technologies could be implemented and meaningfully tested.

Semantic Description of Crowdsourcing Data and Workflows. Data obtained through crowdsourcing should be represented together with relevant metadata such as: the method used to derive the data; the details of the crowdcontributors; the aggregation method used etc. Such metadata allows for the correct interpretation and exchange of the crowdsourced data. While some vocabularies have been proposed to represent crowdsourced data (e. g., I. Celino's Human Computation Ontology²), convergence towards standard vocabularies for this purpose is an important future step. Not only the data of crowdsourcing processes but also their internal workflows could be represented using semantic models, drawing upon and adapting, for example, on previous work on Semantic Web Services.

References

- 1 Sabou, M., Bontcheva, K., Scharl, A., and Föls, M. (2013). Games with a Purpose or Mechanised Labour?: A Comparative Study. In Proc. of the 13th International Conference on Knowledge Management and Knowledge Technologies, i-Know'13, pages 1–8. ACM.

² <http://swa.cefriel.it/ontologies/hc>

- 2 Sabou, M., Scharl, A., and Föls, M. (2013). Crowdsourced Knowledge Acquisition: Towards Hybrid-genre Workflows. *International Journal of Semantic Web and Information Systems*, 9(3):14–41.

3.15 Crowdsourcing for Evaluation and Semantic Annotation

Harald Sack (Hasso-Plattner-Institut – Potsdam, DE)

License  Creative Commons BY 3.0 Unported license
© Harald Sack

Evaluation

In general for several evaluation tasks related with Semantic Web applications crowdsourcing is one viable way to achieve a sufficient number of human evaluators for qualitative evaluation. We have already applied crowdsourcing for the evaluation of the following tasks: Fact Ranking [1], Semantic Search, Exploratory Search [2], and Content-based Recommender Systems. All 4 tasks have in common that there is no unique and generally accepted best solution or gold standard available. The most important fact for a given context, the best recommendation based on an example, the best new direction for exploratory search or semantic search. Most times the question of what is the “best result” lies in the eye of the beholder and is dependent on the personal context of the person you ask. Therefore, to create a generally accepted gold standard or evaluation of achieved results, a large number of evaluators must state their opinion about what is best and/or how to rank achieved results. One of the challenges is of course the selection of a representative sample of tasks to be evaluated by a representative (most desirably unbiased in any sense) sample of evaluators. Moreover, for evaluation it is also very important that the evaluators are not cheating. Therefore, an emphasis has to be put on the detection of any kind of fraud within the crowdsourced evaluation task. Also care has to be taken with the decision how to create the task to be solved by the user. It should not be obvious for the user how to influence the outcome of the overall evaluation task in any sense. On the other hand, the task must not be boring to be able to attract a sufficient number of users. Likewise it is also critical that the users possess sufficient expertise to be able to solve the evaluation task in a meaningful way.

Semantic Annotation

A classical task to be solved via crowdsourcing is the provision of (semantic) annotations of any kind of document. This ranges from annotating text with the most suitable keywords, categories, or semantic entities, to the annotation of multimedia documents, such as e.g. images, audio, or video. This is a classical crowdsourcing scenario when dealing with simple textual annotations. With semantic annotations in the sense that predefined semantic categories or even a larger number of potentially suitable semantic entities must be selected by the user, special care has to be taken when designing the user interface in an efficient way. Annotation should be possible in a simple and comfortable way. One example to solve this issue is the provision of auto-suggestion services [3]. In the same way as autocompletion of as e.g. query terms for search engines, autosuggestion takes into account the current user (text) input and suggests potentially suitable semantic entities that fit best to the user’s input. If large knowledge bases are used for this task, as e.g. DBpedia, then also a large number of entities might be suggested due to ambiguities and partial matches. Thus, also for auto-suggestion, it is necessary to rank the suggested entities in a way that the entity which

is most likely to be selected by the user is presented first. This is a problem similar to the previously mentioned fact ranking. For image annotation and esp. for time-based (region-based) video annotation, the design and implementation of an efficient user interface is still an important and not completely solved task. For both tasks, evaluation as well as semantic annotation the design of a game-based approach to collect or to produce the necessary user provided data, seems to be a promising way to attract a large number of users [3, 4]. On the other hand, this approach is also rather expensive in terms of developing time and costs. Thus, it must be decided whether to spend resources for game development or as micro payments, as e. g. in mechanical Turk applications.

References

- 1 A. Thalhammer, M. Knuth, and H. Sack: Evaluating entity summarization using a game-based ground truth, in *The Semantic Web – ISWC 2012* (P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, eds.), *Lecture Notes in Computer Science*, vol. 7650, pp. 350–361, Springer Berlin Heidelberg (2012)
- 2 J. Waitelonis, H. Sack: Towards exploratory video search using linked data, *Multimedia Tools and Applications*, Volume 59, Number 2 (2012), 645-672, DOI: 10.1007/s11042-011-0733-1, Springer Netherlands, 2012.
- 3 J. Osterhoff, J. Waitelonis, H. Sack: Widen the Peepholes! Entity-Based Auto-Suggestion as a rich and yet immediate Starting Point for Exploratory Search, 2. Workshop Interaktion und Visualisierung im Daten-Web (IVDW 2012)
- 4 J. Waitelonis, N. Ludwig, M. Knuth, H. Sack: Whoknows? – Evaluating Linked Data Heuristics with a Quiz that cleans up DBpedia. *International Journal of Interactive Technology and Smart Education (ITSE)*, Emerald Group, Bingley (UK), Vol. 8, 2011 (3)
- 5 L. Wolf, M. Knuth, J. Osterhoff, H. Sack: RISQ! Renowned Individuals Semantic Quiz – A Jeopardy like Quiz Game for Ranking Facts, 7th Int. Conference on Semantic Systems, *ACM Int. Conf. Proc. Series*, ACM Inc, (i-Semantics 2011), Graz (Austria), Sep. 7–9, 2011, pp. 71–78

3.16 Who and How Should Be Involved in Crowdsourced Data Interlinking?

Cristina Sarasua (Universität Koblenz-Landau, DE)

License  Creative Commons BY 3.0 Unported license
© Cristina Sarasua

The heterogeneity of independently curated data sets on the Web of Data makes the data interlinking process challenging for automatic techniques. With this position statement I would like to describe the topics that arose while researching how to integrate microtask crowdsourcing in the process of automatic RDF data interlinking, and which may be relevant in other knowledge-intensive tasks. For example, dynamically identifying the knowledge of particular crowd workers, taking decisions accordingly and analysing the different kinds of interaction between crowd workers and automatic techniques.

Semantic Web Task with Crowdsourcing

The Semantic Web task where I used crowdsourcing is data interlinking, the process of discovering links between data of different RDF data sets. I used microtask crowdsourcing

as a mechanism to involve humans in the process because they are still crucial to support purely automatic interlinking technology in 1) identifying the sources to be connected, 2) training active learning interlinking algorithms and 3) post-processing the outcome of the automatic interlinking techniques. Microtask crowdsourcing provides a cost-effective and scalable alternative to having expert users in the process, and automates a task that is otherwise often not accomplished systematically. My research focuses on the application of microtask crowdsourcing to the specific scenarios of ontology alignment (i. e. mappings between vocabulary terms) [2] and instance data interlinking (i. e. links that show equivalence or any other domain-specific relationship between particular individuals). I encountered several challenges that I would like to discuss: first, I experienced that crowdsourcing a task like data interlinking in domain-specific scenarios (e.g. research data and publications of the social sciences) requires crowd workers to be further instructed both in the task and the domain. A second challenge was to accurately identify how suitable the knowledge of a worker is for processing a particular link. Third, the a priori assessment of the value that the crowd can add to the automatic interlinking of a particular pair of data sets. In my opinion, implementing a mechanism to attract and redirect appropriate crowd workers to available microtasks in online marketplaces could optimise crowdsourced work.

Including Semantics in a Crowdsourcing Task

A crowdsourcing task where I see that semantics might help is the recruitment and selection process in microtask crowdsourcing. A machine-readable and interoperable description of the different crowdsourcing agents (i. e. crowd workers and requesters) and their work experience, can promote the recognition for work across the different microtask crowdsourcing platforms. I recently proposed the use of an ontology-based Crowd Work CV [1], to provide detailed descriptions analogously to traditional CVs. This could indirectly help in quality assurance in for example, crowdsourced interlinking microtasks. Another task of the crowdsourcing process where introducing semantics could obviously be useful is in publishing the results. Some crowdsourcing platforms already enable requesters to publish crowd-generated data. If such data was automatically annotated with existing vocabularies (e.g. with the Ontology for Media Resources) and offered as Linked Open Data, it could be more easily consumed and integrated by third-party applications. The challenges that I identify in these areas are: first, the definition of a common understanding of the Crowd Work CV model, which satisfies the requirements of all microtask platforms and can be easily integrated with other business-related information. Second, the tradeoff between ranking crowd workers based on their CVs and giving them the freedom to work on what they are interested in. Third, the adoption of Semantic Web standards by the crowdsourcing platforms.

References

- 1 Sarasua, C., and Thimm, M. 2013. Microtask available, send us your cv! In Proceedings of the International Workshop on Crowd Work and Human Computation (CrowdWork 2013), co-located with Social Computing and its Applications (SCA2013).
- 2 Sarasua, C.; Simperl, E.; and Noy, N. F. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. In Proceedings of the 11th International Semantic Web Conference (ISWC2012).

3.17 Linking Implicit with Explicit Semantics: An Initial Position Statement

Markus Strohmaier (GESIS & Universität Koblenz-Landau, DE)

License  Creative Commons BY 3.0 Unported license
© Markus Strohmaier

Linking implicit with explicit semantics has been one of the original challenges for the Semantic Web: How can we semantically annotate web pages such that both machines and humans are augmented in exploration, navigation and cognitive tasks such as understanding? Much progress has been made with regard to for example linking data sources (e.g. via Linked Data), annotating HTML pages (via e.g. Schema.org) or annotating scientific literature (Bioannotator). Yet these attempts can only be seen as a first step towards a more comprehensive and more systematically interwoven Semantic Web that seamlessly integrates implicit semantics with explicit semantic representations irrespective of type, format or medium. For example, algorithmically annotating images with adequate and valid semantic descriptors still represents a major challenge. Semantically annotating short texts or the novel language (emojicons, slang, hashtags, etc) that is emerging on social media such as Twitter or Facebook represents another challenge that is far from being solved by the current state of semantic and/or natural language understanding methods and techniques. Finally, assigning accurate semantic descriptors to academic or other domain-specific textual resources that require deep background knowledge for proper understanding represents another example of a challenge that has not yet been met.

At the same time crowdsourcing has emerged as an interesting alternative solution to problems that can not be solved with algorithmic approaches alone. Human judgements organized in micro workflows, and augmented with algorithmic approaches for optimization and quality control have the potential to expand our arsenal of algorithms with a flexible, on-demand oracle that could help address some of the fundamental challenges for the Semantic Web. Understanding and managing the potential of crowdsourcing for linking implicit with explicit semantics should represent a pressing challenge for Semantic Web research and the Semantic Web community at large. This needs to include tackling questions related to the design of proper incentive structures, the development of adequate approaches to quality assurance and to novel approaches to evaluation.

3.18 Opinions and Aims in Participatory Sensing

Gerd Stumme (Universität Kassel, DE)

License  Creative Commons BY 3.0 Unported license
© Gerd Stumme

One of the imminent societal challenges is climate change. Both the avoidance of further climatical changes as well as adaptation to them requires significant changes of our societies and economies and of our individual and collective life styles. The enforcement of novel policies may be triggered by a grassroot approach, with a key contribution from information and communication technology. Nowadays low-cost sensing technologies allow the citizens to directly assess the state of the environment; social networking tools allow effective data and opinion collection and real-time information spreading processes. Moreover theoretical and

modeling tools developed by physicists, computer scientists and sociologists allow to analyse, interpret and visualize complex data sets.

The EveryAware project integrates all crucial phases (environmental monitoring, awareness enhancement, behavioural change) in the management of the environment in a unified framework, by creating a new technological platform combining sensing technologies, networking applications and data-processing tools; the Internet and the existing mobile communication networks will provide the infrastructure hosting such platform, allowing its replication in different times and places. Case studies concerning different numbers of participants will test the scalability of the platform, aiming at involving as many citizens as possible thanks to low cost and high usability. The integration of participatory sensing with the monitoring of subjective opinions is novel and crucial, as it exposes the mechanisms by which the local perception of an environmental issue, corroborated by quantitative data, evolves into socially-shared opinions, and how the latter, eventually, drive behavioural changes.

Enabling this level of transparency critically allows an effective communication of desirable environmental strategies to the general public and to institutional agencies.

3.19 Crowdsourcing and the Semantic Web

Tania Tudorache (Stanford University, US)

License  Creative Commons BY 3.0 Unported license
© Tania Tudorache

What are the Semantic Web tasks where you felt you needed crowdsourcing?

The work I am doing is in the context of the collaborative authoring and maintenance of large biomedical ontologies. There are several tasks in which crowdsourcing could be beneficial: (1) Knowledge acquisition – filling out property values for particular ontology entities (e. g., body part associated to a disease from a predefined value set, or synonyms); (2) label translation (e. g., translating medical titles, definitions, synonyms, etc. to different languages); (3) quality assurance, such as, (a) verifying the class-subclass relations in a disease taxonomy, (b) verifying existing property values, (c) verifying that a textual definition corresponds to a formal definition and vice-versa (e. g., the necessary and sufficient conditions of a class description appear as intended in the textual definition); (4) mapping (e. g., mapping between entities in a disease ontology to entities in another medical ontology).

Why? Even though these tasks require some domain knowledge, they are suitable for crowdsourcing as they can be fairly easily sliced into smaller and independent tasks. Some of the tasks may require additional knowledge from the ontology, but this is usually localized and easily extractable.

Challenges. (1) Finding the workers with the appropriate domain knowledge; (2) For task 1, filling out property values from a value set, the challenge is how to present or even filter the values sets in the task, especially if the value sets are larger or hierarchical; (3) expert curation of the crowdsourcing results – once the crowdsourcing results are back, how do the domain experts “vet” them. This is especially important for the knowledge acquisition task of large ontologies, in which high quality results are expected.

What are the crowdsourcing tasks where using semantics might help?

Crowdsourcing could benefit from access to structured and interlinked data, for example: (1) Creating qualifying questions: well established ontologies could provide the source for creating the qualifying questions (e. g., both the taxonomy and property values could be used to generate the questions). (2) Creating “intelligent” and adaptable tasks in an iterative workflow: ontologies and vocabularies can provide the knowledge for generating iteratively enhanced tasks. The results of a task together with the information from the ontology will be used to generate more specific tasks by filtering or pruning out invalid knowledge (e. g., in a combinatorial problem, excluding a taxonomy branch, if the parent was excluded). (3) Provide context for a task: some tasks require background knowledge, which can be more easily obtained from the Linked Data cloud, as the information is structured, rather than from an unstructured source. (4) Quality assurance: similar to task 1, the information from established ontologies and vocabularies could be used to create trick questions that quality workers are expected to respond in conformance with the ontology.

Why? Information on the Semantic Web and Linked Data cloud provide a structured access to data that makes it easier to use in a programmatic way. SW data also provides well-agreed upon background knowledge in form of ontologies and vocabularies that can be useful for different tasks (see above).

Challenges. (1) Identifying the “right” ontologies and vocabularies to be used for a certain domain and task. (2) Identifying the minimum context for a task that can be extracted from the Linked Data cloud, which provides the most useful information to a worker. (3) Providing an easy to use interface (e. g., in forms of software libraries) for working with Semantic Web data (access to linked data or ontologies) for different parts of the crowdsourcing workflow.

3.20 The Role of Crowdsourcing and Semantic Web for Consumable APIs

Maja Vukovic (IBM TJ Watson Research Center – Yorktown Heights, US)

License  Creative Commons BY 3.0 Unported license
© Maja Vukovic

Loose coupling and scalability characterize RESTful (REST) architectures. Simplicity of REST Application Programming Interfaces (APIs) has resulted in rapid development of highly consumable services, through power of reuse [1, 2]. This opens up a significant opportunity to create new service capabilities based on existing REST APIs. This cocreation results in diffused networks of API providers and consumers.

Most commonly APIs are described in terms of endpoints, data formats, and protocols to ease the consumption or even composition of APIs. Researchers are considering novel semantic models and graph-based methods for API discovery [3]. For example, API graph [4] aims to enrich the API descriptions semantically to facilitate not only consumption, but also the provisioning of APIs and the evolution of the API ecosystem. It provides API consumers with information about valuable API composition. It lets API providers benefit from competitive analysis with other APIs. Finally, it enables ecosystem providers to identify gaps in API capabilities and their demand. Early models about incorporating semantics into API networks have emerged [4]. Moreover, this offers opportunities for automatically testing if the data exposed by the APIs follow the principles of linked data [5] so that they can be

interlinked and become more useful. As a result, building a better semantic understanding of each API, its use, and attributes can minimize some of the consumability risks. And the key role here is the one of expert crowd, developers, consumers, and integrators.

Crowd is commonly taking the role of a sensor, actuator and controller in a variety of computing [6]. My interest is in identifying the best way to enable crowdsourcing to enrich the API semantic descriptions, which will lead to more consumable services. For example, can we engage developers, by providing crowdsourcing tasks to label and validate their APIs (within the IDE)? How can consumers participate in this process more effectively? What is the right “task size” and incentive to provide in this domain?

References

- 1 Pautasso C., O. Zimmermann, and, F. Leymann. Restful web services vs. big web services: making the right architectural decision. Proceedings of the 17th international conference on World Wide Web. ACM, 2008.
- 2 DuVander A., “9,000 APIs: Mobile Gets Serious,” ProgrammableWeb, <http://blog.programmableweb.com/2013/04/30/9000-apis-mobile-gets-serious/>, 2013.
- 3 Dojchinovski M., J. Kuchar, T. Vitvar, and M. Zaremba. Personalised Graph-based Selection of Web APIs. In Proceedings of the 11th International Semantic Web Conference (ISWC). Lecture Notes in Computer Science, no. 7649. Springer Berlin Heidelberg, 2012.
- 4 Wittern E., J. Laredo, M. Vukovic, V. Muthusamy, A. Slominski. A Graph-based Data Model for API Ecosystem Insights. IEEE International Conference on Web Services. 2014.
- 5 Bizer C., T. Heath, and T. Berners-Lee. Linked data-the story so far. International Journal on Semantic Web and Information Systems (2009): 1–22.
- 6 Vukovic, Maja, R. Das, S. Kumara. From sensing to controlling: the state of the art in ubiquitous crowdsourcing. International Journal of Communication Networks and Distributed Systems. Volume 11, Number 1. 2013. Inderscience Publishers.

3.21 Training Systems with Crowd Truth

Christopher A. Welty (IBM TJ Watson Research Center – Hawthorne, US)

License © Creative Commons BY 3.0 Unported license
© Christopher A. Welty

Crowdsourcing is often used to gather annotated data for training and evaluating computational systems that perform semantic interpretation, such as natural language processing. Crowd workers are asked to perform the same semantic interpretation as the computational system to establish a “ground truth”. I will discuss the use of Lora Aroyo’s Crowd Truth paradigm to collect human annotated data from the crowd for training a machine learning component that performs the NLP task of relation extraction, with examples and experimental results. I will argue that our results support the hypothesis that semantic technologies reliance on “single truth” styles of semantics are flawed and need to be revisited.

3.22 Merging Contexts

Marco Zamarian (University of Trento, IT)

License  Creative Commons BY 3.0 Unported license
© Marco Zamarian

It seems obvious that we are still far from producing a reliable, automated, method for generating even relatively simple semantic content in cases where context matters. Here, human input is clearly needed and crowdsourcing technologies can help tap this input. A fairly good example might be the case of the development of a common understanding between separated communities sharing a common problem but not a common language. In these cases the exchange of information is typically flawless within each community but becomes almost impossible between the communities. The problem is exacerbated when the two communities do not share a similar level of understanding (i. e. in terms of context) of the problem, and thus basic vocabularies do not overlap in semantic terms. This is the case of rare diseases in medicine, where often there are two (lets consider the simple case of an established syndrome, and not the more complex case of the process of discovery of a new one) clearly defined communities, the experts conducting research on the syndrome on one side, and the patients on the other. The first community shapes up by a process of mutual, scientific recognition, and develops a highly contextualized vocabulary to exchange information on the disease, creating the conditions to, at least in principle, solve the semantic problem. On the other hand, the second community faces a much more challenging situation. Its members are more isolated; they do not have the basic knowledge to access relevant information (even a simple online search by keywords might be beyond their knowledge); in almost all cases they need an interface (namely their family practice doctors) to get access to the first community; their concerns (and thus the kind of information they might be searching for) are totally different from the ones occupying the community of experts. Links and fruitful exchanges of information in these cases occur when common, shared events are organized, however events of this kind are infrequent. Thus, building a platform allowing these separated communities to share information, on the one hand seems like an almost ideal setting for the application of semantic technologies, however spurring the two communities to engage in the production of contents that can be fruitfully exchanged asks for a thoughtful approach. Crowdsourcing, per se, can be thought of as an exercise in developing complex sets of incentives that spur people to engage in activities that can easily become very complex, to the point of being beyond the reach of each contributor. In this specific case, which one can claim to be a typical example of classes of situations where developing vocabularies spanning different communities or different contexts is the goal, the challenges are many and they can be separated into at least two broad categories. First, it is obvious that we are talking about separable communities of potential crowdworkers: devising ways to filter and separate access is non obvious and it could involve ambiguities (what community does a common physician belong to?). Second, there is the need to create separate incentive schemes to participate for members of the two communities. Obviously the experts would be able to contribute more, and more in-depth to the semantic tasks, but are less keen to do so (creating a shared representation is less useful for them, and sharing could potentially undermine their status in their community). The opposite would be true for non-experts.

Participants

- Maribel Acosta
KIT – Karlsruher Institut für
Technologie, DE
- Sofia Angeletou
BBC – London, GB
- Lora Aroyo
Free Univ. of Amsterdam, NL
- Abraham Bernstein
Universität Zürich, CH
- Irene Celino
CEFRIEL – Milano, IT
- Philippe Cudré-Mauroux
University of Fribourg, CH
- Roberta Cuel
University of Trento, IT
- Gianluca Demartini
University of Fribourg, CH
- Michael Feldman
Universität Zürich, CH
- Yolanda Gil
University of Southern California
– Marina del Rey, US
- Carole Goble
University of Manchester, GB
- Robert Kern
IBM Deutschland – Böblingen,
DE
- Jan Marco Leimeister
Universität Kassel, DE &
Universität St. Gallen, CH
- Atsuyuki Morishima
University of Tsukuba, JP
- Natasha Noy
Google Inc. –
Mountain View, US
- Valentina Presutti
CNR – Rome, IT
- Marta Sabou
MODUL Universität Wien, AT
- Harald Sack
Hasso-Plattner-Institut –
Potsdam, DE
- Cristina Sarasua
Universität Koblenz-Landau, DE
- Elena Simperl
University of Southampton, GB
- Markus Strohmaier
Universität Koblenz-Landau, DE
- Gerd Stumme
Universität Kassel, DE
- Tania Tudorache
Stanford University, US
- Maja Vukovic
IBM TJ Watson Research Center
– Yorktown Heights, US
- Christopher A. Welty
IBM TJ Watson Research Center
– Hawthorne, US
- Marco Zamarian
University of Trento, IT



Information-Centric Networking 3

Edited by

Dirk Kutscher¹, Taekyoung Kwon², and Ignacio Solis³

1 NEC Laboratories Europe – Heidelberg, DE, dirk.kutscher@neclab.eu

2 Seoul Nat. University, KR, tkkwon98@gmail.com

3 PARC – Palo Alto, US, Ignacio.Solis@parc.com

Abstract

This report documents the presentations and discussions of the 3rd Dagstuhl seminar on Information-Centric Networks. This seminar was focused on the deployment and scalability of ICNs. An overview of various ICN projects was used as a starting point for discussions. Participants provided a set of starting questions to cover with the rest of the group. The seminar increased the awareness on the state of the art in ICN research. Various topics on deployment and scalability were discussed. The opinions and comments presented here came directly from the notes taken at the seminar.

Seminar July 13–16, 2014 – <http://www.dagstuhl.de/14291>

1998 ACM Subject Classification C.2.1 Network Architecture and Design

Keywords and phrases Information-Centric, Content-Centric, Name-Based, Content-Based, Networks

Digital Object Identifier 10.4230/DagRep.4.7.52

1 Executive Summary

Ignacio Solis

License © Creative Commons BY 3.0 Unported license
© Ignacio Solis

Information Centric Networks (ICN) has been a growing area of research in the past few years. The Dagstuhl ICN Seminar series has played a central role in forming the research community. The first seminar, Dagstuhl Seminar 10492, was the meeting point of the various ICN projects across the world; both from the academic perspective as well as the commercial perspective.

The community created at this event continued interacting. It was not long before the members created a set of academic workshops at the most important networking conferences; SIGCOMM, INFOCOM, etc. Following the success of the second Dagstuhl Seminar (12361), the community continued to coalesce and founded the ICNRG. The ICNRG, Information Centric Networking Research Group, was formed at the IRTF to evaluate the technology and to create a forum for companies discuss possible standardization efforts.

With the third iteration of this Seminar we've attempted to bring together the academic and commercial community together once more to discuss the state of the art in ICN. Specifically, we've focused on scalability and deployment. First, what are the problems we face in terms of scaling ICN. Are there technical limitations or political limitations. Second, what are the roadblocks in the path towards deployment. Since there will be no overnight switch, the technology must be deployed in controlled environments where interoperability can be slowly achieved.



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

Information-Centric Networking 3, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 52–61

Editors: Dirk Kutscher, Taekyoung Kwon, and Ignacio Solis



Dagstuhl Reports

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

2 Table of Contents

Executive Summary	
<i>Ignacio Solis</i>	52
Overview of Talks	
ENCODERS	
<i>Ashish Gehani & Minyoung Kim</i>	54
CASCADE	
<i>Vikas Kawadia</i>	54
Green ICN	
<i>Tohru Asami</i>	54
NetInf Update	
<i>Börje Ohlmann</i>	55
CCN 1.0 Update	
<i>Ignacio Solis</i>	55
CCN Lite	
<i>Christian Tschudin</i>	56
Discussion Starters	57
Discussion Groups	58
Participants	61

3 Overview of Talks

3.1 ENCODERS

Ashish Gehani (SRI – Menlo Park, US), Minyoung Kim (SRI – Menlo Park, US)

License  Creative Commons BY 3.0 Unported license
© Ashish Gehani & Minyoung Kim

ENCODERS is part of the DARPA CBMEN program. CBMEN (Content-Based Mobile Edge Networking) tries to build an Information Centric approach for soldiers to network in the field. It focuses on a mixture of ICN and MANET. ENCODERS uses a mixture of push and pull. It incorporates a priority system that is used between the publisher and requester to negotiate queuing and caching resources. The core of the system is based on Haggie. Discussion questions

- What information to be protected how?
- Can the insiders leak enough information so that outsiders can become insiders?
- How do you handle revocation?
- What are the energy consumption for security?

3.2 CASCADE

Vikas Kawadia (BBN Technologies – Cambridge, US)

License  Creative Commons BY 3.0 Unported license
© Vikas Kawadia

CASCADE is part of the DARPA CBMEN program. CBMEN (Content-Based Mobile Edge Networking) tries to build an Information Centric approach for soldiers to network in the field. It focuses on a mixture of ICN and MANET. CASCADE is a content-based architecture based on the notion of communities. Nodes keep data around per community and move the data by duplicating it between communities. It's built on top of TCP and UDP. CASCADE uses a put and get API.

It is similar to a DHT over a MANET where the DHT focuses on a specific region (community). The core notion is that the region is stable. Basically, network links are stable between members. This region can move; for example people on a bus.

Content management is done via machine learning using meta-data. This meta-data informs the system of which data must move and where. There is an assumption of high heterogeneity.

3.3 Green ICN

Tohru Asami (University of Tokyo, JP)

License  Creative Commons BY 3.0 Unported license
© Tohru Asami

Green ICN is a project on using ICN in situations with low amounts of energy. Focusing on disaster situations. Two main scenarios being considered:

- Disaster management (aftermaths of disaster): the green part is essential to make devices last as long as possible after the disaster so to keep critical service running (intervention time may be long) => energy consumption is a priority in this scenario
- Video delivery: how to make video distribution at very large scale

Disaster scenario: After the Japanese earthquake some base stations were working, but the backhaul was not. One of the main goals is to enable communication between mobiles for 3 days (it took 3 days to get the backhaul up again).

The project considers more than just having some communication. For example, in the case of the government wanting to communicate with the citizens not only must there be a way to distribute the message but also a way for users to be able to authenticate the messages.

Notes:

- Feedback from 9/11 shows the amount of communication needed for basic services is small. But getting data from everybody and sharing that is expensive.
- Security and authentication might work against efficiency. It may be better to accept the noise and let the crowdsource information flow.
- Can there be better ways to summarize information? We don't need 10000 reports of something going wrong (or right).
- How do you prioritize traffic between the population and the government?

3.4 NetInf Update

Börje Ohlman (Ericsson Research – Stockholm, SE)

License © Creative Commons BY 3.0 Unported license
© Börje Ohlmann

NetInf prepared a description of their project for live video streaming of the Ski World Cup Falun 2015. Ericsson worked on a “Virtual Arena” platform. It presented data and video streaming via web and other systems. The goal of the project was to see what NetInf technology brings to the table. Can it offer live streaming to Android as part of the virtual arena? This project is in the design stages but the plans are to open source the code. It's unclear what the success metrics for the project are at this point.

3.5 CCN 1.0 Update

Ignacio Solis (Xerox PARC – Palo Alto, US)

License © Creative Commons BY 3.0 Unported license
© Ignacio Solis

PARC has been working on CCN 1.0. It is a complete redesign of CCNx; both the protocol and the code has changed.

The base CCN protocol is the same. An interest message acts as a request for named data. Applications issue interests to retrieve named information. A content packet is a reply to this request. It carries a piece of named data. Names in CCN are a sequence of binary segments (or components). Each content packet (or content object) is named and contains a signature. This signature binds an identity, the name and the payload together into a

content object. When an application receives a content object it can check the validity by checking the signature.

Content Objects are network level elements, they are not the same as application level elements (like files). Large files are divided into chunks. Each chunk forms a content object. Chunks can be up to 64K in size but are normally the size of a regular network MTU (1500 bytes). Files are described using manifests. These manifests can be signed like other objects and list the set of content objects that make up the file. This list of content objects is created using self-certified names (hash based names of the content objects). This means that the content objects do not require a direct signature since they are implicitly signed from the manifest.

Changes from CCN 0.x to CCN 1.x:

- Packet format is now based on TLVs. 2+2 TLVs (2 bytes for T, 2 bytes for L).
- Packets now have a static header and some optional headers.
- Content Object matching is now done with exact match.
- Interests no longer have selectors.
- Interests can now have payloads.
- Both Interests and Content Objects have a unified format.
- The unified packet format contains the validation information at the end to make it modular.

The CCN 1.x software suite is being coded from scratch. It is written in C and uses modern software development practices. All code is unit tested and documented. It follows a specific coding style. Code complexity is low.

The current design uses a component based transport framework. This framework has various components that are used to instantiate stacks at runtime. Transport stacks can be configured to do different things. Some stacks deliver content in order and reliably. Other stacks might focus on pre-signed content, etc.

A set of APIs will be built on top of the transport framework. Applications will not need to interact with the network via sockets. The current plan is to have a KeyValue Store, a Messaging API and a streaming API. The native API currently used is called Portal. It allows applications to manually construct Interests and Content Objects.

3.6 CCN Lite

Christian Tschudin (Universität Basel, CH)

License  Creative Commons BY 3.0 Unported license
© Christian Tschudin

The goal of CCN Lite is to create a basic CCN forwarder in less than 1000 lines of C code. The current instantiation runs in user space and has basic interoperability with CCN 0.8 using the ccnb packet format. They are working on getting the capability to tunnel alien encoding formats. Code base is now part of the RIOT project (FU Berlin).

Christian is also working on Named Function Networking (NFN). This is a project to do lambda expressions over the network, basically asking the network to perform computation: “Network, please do this and give me the result”.

4 Discussion Starters

We requested participants to submit various discussion starters for the breakout sessions. These discussion starters were presented to the group first to determine the most popular topics and to allow the participants to pick which topic they wanted to discuss.

Deployment and adoption

- Networking today is like making a phone call. With a socket we connect somewhere and then exchange what we want. That's changing, but to what? Should we include ICN names directly in the programming languages?
- People feel proprietary in that they want to know who got their data, and also want to know from whom they got the data, thus motivating the security functions we have today (argument for a connection model?)
- How could ICN, through some short-term adoption, provide value for content providers and users, and thus start deployment?
- We need a killer app – “Killer app” in the sense that it can kill your project: debugging, monitoring, traffic engineering, etc; e.g, ping, and other management tools; this is dissemination critical – we have to get that right!

Architecture scalability

- Scaling the overhead of locating a content in a network; if things move around you need ways to handle this; if you want to find the nearest copy, how do you find it and what control overhead does that result in? how do you synchronize state in different routers?
- What is the role of centralized services (Google) in finding content?
- What is the current feasible line-rate?
- What do we need to do cache-aware routing – architecture for optimizing routing, to find a copy close to the requester; can use DHT or name resolution system to locate near copies.

Domains and traffic

- How can we scale ICN over multiple administrative domains? what protocols are needed, policy mechanisms, etc?
- How can we increase the efficiency of handling multimedia and in particular video using ICN?
- One of the problems with user experience is the dual control loops that don't know about each other; a different way is to get rid of one of them, running the application directly on top of a simpler ICN service; what are the fundamental constraints the encoders have? how would that map to ICN without trying to mimic how we today do it on top of TCP?

5 Discussion Groups

With the discussion topics picked we proceeded to breakout sessions.

Scalability

- Routing systems – what’s the benefit of caching?
- How to utilize the cache in the routing system? One way is to look in your local domain, then forward to a peer with whom you have a bus request, else forward to origin domain.
- Is there a need to have locations and name locations?
- Will the resolution system resemble the current DNS?
- What’s the overhead of interdomain caching?
- Will systems advertise prefixes to business partners?
- NetInf naming could potentially be used – covers cache-aware location names, allows to advertise location and comment.
- What’s the relation of scalability to content names?
- Forwarding plane scalability has been address before, so not a lot of discussion on this.
- Integrating caching and forwarding – Cisco says can support forwarding @ 20Gbps. Haven’t figured how to check cache at the same time at that rate. Infocom paper this year about cached video.
- Verify signature on each packet at each hop seems clearly infeasible. Need to design in a way that gives realistic assumptions.

API

Many participants were at previous ICN workshops, felt we circled around questions that had been circled around before. [Dave Joke: loop detection but not loop suppression.]

Philosophical argument: which “PI” are we talking about? Facing the application or top of the network service. Embedded in the discussion is question of which applications exist, who is responsible for what aspects of managing the content.

Three examples talked through – all have “application” and network.

- Publisher (“Hollywood”) has created a static object (movie). Their view is they want to put it into the cloud and forget about it – network will distribute it. People who are interested in watching it can issue request/interest in the movie and get the movie back. Caching helps efficiency, maybe performance. There is an argument about where is the storage in the network.
- “Dark content” – content that doesn’t already exist when it is requested (e.g., credit card account balance report). Publisher only produces the report when it is asked for – but routing system needs something so the network knows how to route queries for the object. Publisher (bank) gets interest, generates the report.
- traveler @ airport, dumps photos @ airport before getting on plane home, wants local repo to get it to home while traveler is on airport. Again, publisher vanishes after “putting”.

Discussed about incentives and financial considerations. Does the API have to deal with that? That led to discussion about controlling access (movie scenario) and watermarking. Who pays the “custodian” in the photo scenario. What exactly is a first-class service? Is storage a first-class service?

What is supported by every node? What is the “core service”? What is “native” and what is “built in”, but not “native”.

Video

- Current state – HLS/Dash-based over ICN. Decouple producer and consumer. What are the real benefits?
- Scenarios – Netflix (least challenging), live event w/synch, video conference.
- Use case: anticipatory Video streaming – network (cell) operator knows where users are and how many – benefit: control greediness of users; helpful in managing buffer behavior. Modified manifest of HLS (similar to CCN manifest). This is anathema of ICN, but also in line with ICN. Can you do a receiver-based mechanism that provides benefits of ICN?
- Use case: video conference system: can I fast fwd through parts of the meeting that I missed? Time shifting video conference. ICN has in-network caching; utilize cache.
- In some cases (secure multicast) with “forward secrecy” you should not be able to decrypt anything that happened before you joined.
- Use case: infrastructure-less. storage on mobile devices, video clips to share, combine network fabric and database. Baseline is very low; lots of potential for improvement.
- Use case: massive public event (stadium scenario). Compute fns at the edge (fog computing), virtualize network elements... Good example for ICN b/c upload capacity is the bottleneck.
- Scalable video? Microsoft lync video coding ideal for ICN. Caching multicasting gain. Fine grained enough for rate control?

Key takeaways:

- ICN Video is a service of the network
- Allows better composition of different elements
- Would benefit from scalable coding
- Need to understand how ICN arch would affect video quality
- Interactive is hard!
- Synch is important issue
- Strategy layer would be useful [yes indeed]
- Content-dependent meta-info is important

Instrumentation

- What is the meta-architecture / philosophy behind instrumenting ICNs?
- Is there a distinction between inside the network to outside the network. (Different than policy permissions)
- Is there a difference between regular users and managers?
- Are we introducing security issues?
- Do we need more than one naming scheme?
- Are you allowed to talk to the cache system?
- Can users have external policies?
- What’s the equivalent of ping in NDN/CCN; reach a namespace? reach anyone?
- Do we need host names for instrumentation purposes?
- Do you use relative or global names for nodes?
- Management might open attacks. We have to deal with those in the same way as we treat other types of attacks.
- Every element of the system has to be able to act as a provider to make it’s state available.
- What is the topology in this system?
- What’s the effect of multi-path?

Security

- scanning vs interest packets
- redirect questions
- ICMP redirect is different
- general concern re: Internet-based DoS attacks
- constraining interests
- make interest expensive or rate limit
- how to protect PIT?
- out of bound mechanisms (make it easy for IT department to find you)
- banking, time constraints

Participants

- Bengt Ahlgren
Swedish Institute of Computer
Science – Kista, SE
- Tohru Asami
University of Tokyo, JP
- Kenneth L. Calvert
University of Kentucky, US
- Antonio Carzaniga
University of Lugano, CH
- György Dán
KTH Royal Institute of
Technology, SE
- Elwyn Davies
Trinity College Dublin, IE
- Anders Eriksson
Ericsson Res. – Stockholm, SE
- Suyong Eum
NICT – Tokyo, JP
- Kevin R. Fall
Carnegie Mellon University, US
- Xiaoming Fu
Universität Göttingen, DE
- Massimo Gallo
Bell Labs – Nozay, FR
- Ashish Gehani
SRI – Menlo Park, US
- Volker Hilt
Alcatel-Lucent – Stuttgart, DE
- Jussi Kangasharju
University of Helsinki, FI
- Holger Karl
Universität Paderborn, DE
- Vikas Kawadia
BBN Technologies –
Cambridge, US
- Minyoung Kim
SRI – Menlo Park, US
- Dirk Kutscher
NEC Laboratories Europe –
Heidelberg, DE
- Taekyoung Kwon
Seoul National University, KR
- Stefan Lederer
Alpen-Adria Universität &
BITMOVIN – Klagenfurt, AT
- Eiichi Muramoto
Panasonic Corporation –
Yokohama, JP
- Edith Ngai
Uppsala University, SE
- Börje Ohlman
Ericsson Res. – Stockholm, SE
- David Oran
Cisco Systems – San Jose, US
- Craig Partridge
BBN Technologies –
Cambridge, US
- Diego Perino
Bell Labs – Nozay, FR
- Ioannis Psaras
University College London, GB
- Damien Saucez
INRIA Sophia Antipolis –
Méditerranée, FR
- Thomas C. Schmidt
HAW – Hamburg, DE
- Glenn Scott
Xerox PARC – Palo Alto, US
- Jan Seedorf
NEC Laboratories Europe –
Heidelberg, DE
- Ignacio Solis
Xerox PARC – Palo Alto, US
- Christian Tschudin
Universität Basel, CH
- Ersin Uzun
Xerox PARC – Palo Alto, US
- Matthias Wählisch
FU Berlin, DE
- Cedric Westphal
Huawei Technologies – Santa
Clara, US
- George Xylomenos
Athens University of Economics
and Business, GR



Report from Dagstuhl Seminar 14292

Network Attack Detection and Defense: Securing Industrial Control Systems for Critical Infrastructures

Edited by

Marc Dacier¹, Frank Kargl², Hartmut König³, and Alfonso Valdes⁴

- 1 Qatar Computing Research Institute (QCRI), Qatar
- 2 Universität Ulm, DE, frank.kargl@uni-ulm.de
- 3 BTU Cottbus-Senftenberg, DE, koenig@informatik.tu-cottbus.de
- 4 University of Illinois – Urbana, US, avaldes@illinois.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14292 “Network Attack Detection and Defense: Securing Industrial Control Systems for Critical Infrastructures”.

The main objective of the seminar was to discuss new approaches and ideas for securing industrial control systems. It is the sequel of several previous Dagstuhl seminars: (1) the series “Network Attack Detection and Defense” held in 2008 and 2012, and (2) the Dagstuhl seminar “Securing Critical Infrastructures from Targeted Attacks”, held in 2012. At the seminar, which brought together members from academia and industry, appropriate methods for detecting attacks on industrial control systems (ICSs) and for limiting the impact on the physical components were considered. A central question was whether and how reactive security mechanisms can be made more ICS- and process-aware. To some extent it seems possible to adopt existing security approaches from other areas (e. g., conventional networks, embedded systems, or sensor networks). The main question is whether adopting these approaches is sufficient to reach the desired level of security for ICSs. Detecting attacks to the physical components and appropriate reactions to attacks are new aspects that need to be considered as well. The main result of the seminar is a list of recommendations for future directions in ICS security that is presented in this report.

Seminar July 13–16, 2014 – <http://www.dagstuhl.de/14292>

1998 ACM Subject Classification K.6.5 Security and Protection, C.2.0 General, J.7 Computers in Other Systems

Keywords and phrases Security, Intrusion Detection, Critical Infrastructures, Industrial Control Systems, SCADA, Vulnerability Analysis, Malware Assessment, Attack Response and Countermeasures

Digital Object Identifier 10.4230/DagRep.4.7.62

Edited in cooperation with Rens van der Heijden



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

Network Attack Detection and Defense: Securing Industrial Control Systems for Critical Infrastructures, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 62–79

Editors: Marc Dacier, Frank Kargl, Hartmut König, and Alfonso Valdes



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

1 Executive Summary

Marc Dacier

Rens van der Heijden

Frank Kargl

Hartmut König

Alfonso Valdes

License © Creative Commons BY 3.0 Unported license
© Marc Dacier, Rens van der Heijden, Frank Kargl, Hartmut König, and
Alfonso Valdes

From July 13–16, 2014, more than 30 researchers from the domain of critical infrastructure security met at Schloss Dagstuhl to discuss the current state of security in industrial control systems.

Recent years have highlighted the fact that security precautions of information and communication technology (ICT) in many critical infrastructures are clearly insufficient, especially if considering targeted attacks carried out by resourceful and motivated individuals or organizations. This is especially true for many industrial control systems (ICS) that control vital processes in many areas of industry that are relying to an ever-larger extent on ICT for monitoring and control in a semi or fully automated way. Causing ICT systems in industrial control systems to malfunction can cause huge economic damages or even endanger human lives. The Stuxnet malware that actually damaged around 1000 Uranium enrichment centrifuges in the Iranian enrichment facility in Natanz is the most well-known reported example of an ICT attack impacting ICS.

This situation led to increased efforts in research which also resulted in a number of Dagstuhl seminars related to this topic of which this seminar is a follow-up event, namely two Dagstuhl seminars on “Network Attack Detection and Defense” in 2008 and 2012 and one on “Securing Critical Infrastructures from Targeted Attacks” held in 2012. The main objective of our this latest seminar was to discuss new approaches and ideas on how to detect attacks on industrial control systems and how to limit the impact on the physical components. This is closely coupled to the question of whether and how reactive security mechanisms like Intrusion Detection Systems (IDS) can be made more ICS- and process-aware. To some extent it seems possible to adopt existing security approaches from other areas (e. g., conventional networks, embedded systems, or sensor networks) and one of the questions is whether adopting these approaches is enough to reach the desired security level in the specific domain of industrial control systems, or if approaches specifically tailored for ICS or even single installations provide additional benefits.

The seminar brought together junior and senior experts from both industry and academia, covering different scenarios including electrical grids, but also many other control systems like chemical plants and dike or train control systems. Apart from the detection and prevention of attacks by both security and safety mechanisms, there was an extensive discussion on whether or not such systems should be coupled more strongly from a security perspective. It was also argued that there exists a very diverse space of application domains, many of which have not yet been subject to much study by security researchers, for various reasons. Many of these discussions were triggered by plenary or short talks, covering topics from the state of the art in ICS security, forensics in ICS, security assessments, and the new application domain of flood management.

Apart from talks and subsequent discussions, a number of working groups were organized during the seminar, intended to address specific issues in the field. In total, there were four

working groups, each of which provided a summary of their results included in this report. The first was on forensics, discussing how attacks can be detected and analyzed after the fact. A second working group addressed the issue of security and risk management, analyzing why existing IT security approaches do not work for ICS and discussing potential improvements. Industry 4.0 and the wide range of new and non-classical ICS use cases was the topic of a third working group, which discussed the new security challenges arising from these emerging research topics. Finally, there was a working group on the detection of cyber-physical attacks; a core question here were advantages and disadvantages of process-aware intrusion detection mechanisms. The group also discussed the interaction between intrusion detection, intrusion response, and security management.

Based on the talks, discussions and working groups, the Dagstuhl seminar was closed with a final plenary discussion which summarized again the results from the working groups and led to a compilation of a list of open issues that participants consider necessary to be addressed. Those issues partly overlap with the list of open issues identified in the seminar proposal but also uncovered many new challenges that may become highly relevant research topics and may lead to a new agenda for future research. Those issues are discussed at the end of this report.

2 Table of Contents

Executive Summary

Marc Dacier, Rens van der Heijden, Frank Kargl, Hartmut König, and Alfonso Valdes 63

Plenary Talks

Examples of Cyber-attacks on SCADA Systems for the Electrical Grid and their Consequences
Gunnar Björkman 66

Towards Resilient Control of Critical Infrastructures
Alvaro Cárdenas Mora 66

Security of Train Control Systems
Stefan Katzenbeisser 67

Short Talks

Performing Forensic Investigations of Industrial Control Systems
Heiko Patzlaff 67

Automatic Analysis of Unknown Network Protocols
Konrad Rieck 67

Are you threatening my Hazards?
Marina Krotofil 68

Secure Our Safety: Building Cyber Security for Flood Management
Dina Hadziosmanovic 68

Security Assessment and Intrusion Detection for Industrial Control Systems
René Rietz and Andreas Paul 69

ICS Security – Challenges, State of the Art and Requirements
Ulrich Flegel 69

Working Groups

Security Consequences and Quantitative Risk Analysis
Stefan Katzenbeisser and Rens van der Heijden 69

Industry 4.0
Nils Aschenbruck and Alfonso Valdes 71

(Real-time) Detection of CPS Attacks
Alfonso Valdes and Rens van der Heijden 73

Cyberforensics
Heiko Patzlaff and Stephan Kleber 75

Open Issues 77

Participants 79

3 Plenary Talks

3.1 Examples of Cyber-attacks on SCADA Systems for the Electrical Grid and their Consequences

Gunnar Björkman (ABB – Mannheim, DE)

License © Creative Commons BY 3.0 Unported license
© Gunnar Björkman

Main reference EU FP7 Project VIKING

URL <http://www.kth.se/ees/omskolan/organisation/avdelningar/ics/research/cc/v>

The presentation gave a number of examples of possible cyber-attacks on SCADA system used for the supervision and control of the electrical grid. It should be noted that these were examples of possible cyber-attacks, not a description of real incidents although some the attacks were inspired from real events like Stuxnet. The included scenarios have been verified by major transmission grid operators and found to be realistic in the meaning that they would be possible to conduct and that the reported consequences are realistic. The attack scenarios are a subset of the story boards developed in the FP7 project VIKING which ended in 2011. The scenarios described in the presentation had been selected to represent attacks of different categories like Denial of Service, Social Engineering, lack of Security Training, etc. and to demonstrate both major and minor consequences from the triggered black-outs in the society. For each scenario the probability for attack success calculated applying the CySeMoL method on the given ICS configuration was reported. The Cyber Security Modeling Language (CySeMoL) is also a result from the VIKING project. In addition to the scenario description and the attack success, two types of societal consequences were described in the presentation. One was a monetary cost which illustrates the lost Gross Domestic Product caused by the blackout and the other a logarithmic value calculated from impacted number of people and black-out duration that indicates non-economic consequences in the society. The latter measure closely resembles the Richter scale and gives an intuitive feeling for the seriousness of an attack on the power grid.

3.2 Towards Resilient Control of Critical Infrastructures

Alvaro Cárdenas Mora (University of Texas at Dallas, US)

License © Creative Commons BY 3.0 Unported license
© Alvaro Cárdenas Mora

The protection of industrial control systems is usually achieved via a series of safety, fault-tolerant, and robust control mechanisms. These solutions were design under the assumption of a non-strategic adversary and targeted mostly random failures.

In this talk we discussed possible new directions on how to extend these previous solutions and adapt them to be resilient to strategic adversaries that will try to evade anomaly detection schemes while at the same time maximize their negative impacts. The talk also described a generic approach to identify the vulnerability of systems at the physical and control layer via controllability and observability concepts, and then proposed resilient control algorithms that can survive their mission critical objectives even when faced with successful attacks partially compromising the system. The speaker showed use-cases in industrial process control of anaerobic and chemical reactors, and in frequency control as well as demand-response control in the power grid.

3.3 Security of Train Control Systems

Stefan Katzenbeisser (TU Darmstadt, DE)

License © Creative Commons BY 3.0 Unported license
© Stefan Katzenbeisser

In the talk Stefan Katzenbeisser gave an overview of the technical systems providing safe train operations in Germany. In particular, he discussed safety and security problems raised by these deployed systems. Furthermore, he also reported on an ongoing work that attempts to provide IT security recommendations for the railway industry, ranging from risk analysis over security aware design up to security management aspects. This set of recommendations is currently proposed as a standard in the DIN German Institute for Standardization. Finally, open research problems in this domain were surveyed.

4 Short Talks

4.1 Performing Forensic Investigations of Industrial Control Systems

Heiko Patzlaff (Siemens – München, DE)

License © Creative Commons BY 3.0 Unported license
© Heiko Patzlaff

Performing computer forensic investigations in industrial control systems (ICS) presents various challenges. Starting with Stuxnet, the general lack of tools and procedures for analyzing security incidents in industrial settings has become apparent. The talk gave an introduction to ICS. It looked at the challenges that arise when one needs to analyse security incidents in industrial products. In particular, what data is available in these systems? How do you acquire this data? How do you transfer and analyse it? And what conclusions can you draw from it? The talk provided some examples of real world cases. And it presented results from a research project aimed at developing tools and approaches for performing computer forensic investigations in ICS that led to the development of a new forensic platform for the investigation of such incidents.

4.2 Automatic Analysis of Unknown Network Protocols

Konrad Rieck (Universität Göttingen, DE)

License © Creative Commons BY 3.0 Unported license
© Konrad Rieck

Joint work of Rieck, Konrad; Krueger, Tammo; Gascon, Hugo; Krämer, Nicole

Main reference T. Krueger, H. Gascon, N. Krämer, K. Rieck, “Learning stateful models for network honeypots,” in Proc. of the 5th ACM Workshop on Security and Artificial Intelligence (AISec’12), pp. 37–48, ACM, 2012.

URL <http://dx.doi.org/10.1145/2381896.2381904>

Proprietary protocols in ICS are a major hurdle for traffic analysis and intrusion detection. Without detailed knowledge of message formats and protocol states, there is little chance that attacks can be spotted in the traffic of these protocols. This talk provided an overview of techniques for automatic protocol reverse engineering. Different approaches for inferring message formats and protocol state machines from network traffic are presented and possible applications for securing ICS were outlined.

4.3 Are you threatening my Hazards?

Marina Krotofil (Hamburg University of Technology, DE)

License © Creative Commons BY 3.0 Unported license
© Marina Krotofil

Joint work of Krotofil, Marina; Larsen, Jason

Main reference M. Krotofil, J. Larsen, “Are You Threatening my Hazards?,” in Proc. of the 9th Int’l Workshop on Security (IWSEC’14), LNCS, Vol. 8339, pp. 17–32, Springer, 2014.

URL http://dx.doi.org/10.1007/978-3-319-09843-2_2

Cyber-physical systems (CPS) are characterized by an IT infrastructure controlling effects in the physical world. On one hand, embedded computers enable governing of physical applications to achieve desired outcomes. On the other hand, physical systems can be instructed to perform actions that are not intended in the same way. Cyber-attacks on physical systems are correspondingly called cyber-physical attacks. The implications of this class of cyber-attacks (the ability to inflict physical damage) is the main difference between cyber-physical and conventional cyber-attacks.

In the context of CPS, safety systems have the critical function of detecting dangerous or hazardous conditions and taking actions to prevent catastrophic consequences on the users and the environment. Relationship between safety and security is usually considered in the context of dependable computing with a focus on IT or system-design. The limitation of this approach is lacking the view of what is happening in the physical world and whether the system remains safe. In our approach we take physical processes into account and propose a simple mental model of the relationship between security assisting why the understanding the hazards within the systems is crucial to designing its security architecture.

The talk covered such aspects as trustworthiness or veracity or process measurements, interactions of the cyber/physical systems (“unexpected physics”), hidden impact data (“unseen” influence of components on each other), hazard-aware security zoning. The latter one allows to harmonize security and safety lifecycles. The system remains secure if updated often (e.g. patched); the system remains safe if untouched. A granular architecture can be created by tracing specific hazards back through a cyber-physical system matching specific devices and specific data flows caring data related to a specific hazard. Components associated with severed hazards must be protected more vigorously.

4.4 Secure Our Safety: Building Cyber Security for Flood Management

Dina Hadziosmanovic (TU Delft, NL)

License © Creative Commons BY 3.0 Unported license
© Dina Hadziosmanovic

In this short talk Dina Hadziosmanovic presented the key idea of an upcoming project. This project aims at improving the cyber security of critical infrastructures by bridging the gap between safety and security risk management and monitoring. The project uses the context of flood management to provide integrated decision support for incident response related to cyber threats, based on both safety and security science. Firstly, it works on enriching network security monitoring with safety context information. Here, the context consists of static information about the underlying physical process, as well as dynamic information about safety threats (i.e., extreme hydrometeorological conditions). Secondly, the project addresses how to update safety incident response by procedures that include information from security monitoring in assessing the expected effectiveness of responses. The integration of the two innovations will enable adequate responses to flood defence security threats.

4.5 Security Assessment and Intrusion Detection for Industrial Control Systems

René Rietz and Andreas Paul (BTU Cottbus, DE)

License  Creative Commons BY 3.0 Unported license
© René Rietz and Andreas Paul

A targeted-oriented improvement of the IT security of Industrial Control Systems (ICS) requires an initial evaluation of the current level of security. Although relevant industry standards in the ICS security domain (e.g., ISO/IEC 27000 series) have already made reference to the need for a comprehensive security analysis, there is no appropriate approach available, yet. This presentation presented a structured, multi-step process for quantitative security assessment of ICS networks. The process yields resilient values, called security indicators, considering the present network topology as well as the technical configuration of the involved industrial network devices. Security indicators can be used to point out potential security vulnerabilities. They also aim to compare the security level of different systems or infrastructures and simulate the application of various measures for improvement. As a novel measure to enhance the security of ICS, an anomaly-detection-based IDS approach was further presented in this talk.

4.6 ICS Security – Challenges, State of the Art and Requirements

Ulrich Flegel (Infineon Technologies – München, DE)

License  Creative Commons BY 3.0 Unported license
© Ulrich Flegel

Security for industrial control system faces a radically different landscape than security for modern information systems. Although these old acquaintances again raise their ugly faces. The talk characterized ICS and summarized the main challenges that security practitioners face when securing such systems. Following it described the current approaches and pointed out their challenges. Ulrich Flegel then summarized the state of the art of available solutions as a baseline to improve upon and finally provided a set of requirements – or rather limitations – that new solutions need to meet.

5 Working Groups

5.1 Security Consequences and Quantitative Risk Analysis

Moderation: Stefan Katzenbeisser (TU Darmstadt, DE)

Minutes: Rens van der Heijden (Universität Ulm, DE)

License  Creative Commons BY 3.0 Unported license
© Stefan Katzenbeisser and Rens van der Heijden

5.1.1 Problem Statement

The goal of this working group was to discuss how to analyze the risk posed by cyber attacks on industrial control systems, with a focus on the potential consequences of these attacks. The quantification of these consequences should help to improve risk assessment

methodologies, and it will allow the design of better detection and preventive measures. This also allows the inclusion of attacks that rely on the attack prevention or safety systems to cause damage, i. e., by triggering a response from these systems the attacker causes financial damage.

5.1.2 Discussion Topics

The discussion covered several topics with opposing viewpoints, specifically regarding how risk analysis should be performed and how much of the physical processes should be considered. We discussed a separation of an ICS into a cyber and a physical component, with limited interaction between them. It was proposed that the physical component mainly needs to be analyzed for safety issues, which can be analyzed using fault trees and analyses like Failure Mode and Effects Analysis (FMEA). Similarly, in the cyber component, attack trees can be employed, as well as analyses like CORAS.

Another question that was discussed is how the cost of different identified security consequences should be considered in risk assessment. It was found that the cost for the compromise of different assets is significantly different in systems with multiple stakeholders, such as a smart metering scenario. This led to a discussion of whether risk assessments should include a focus on societal cost and how quantitative assessments can be adapted for these purposes.

The working group noted that the interaction between the cyber and the physical component is especially interesting for detecting attacks, and therefore should also be considered during risk analysis. Several related important challenges in risk assessment for ICS can be partially addressed in this way.

One of these challenges is that the amount of possible attacks and the lack of information regarding actual attacks mean that quantifying the risk is a difficult challenge compared to normal risk analysis common for security. This can be compensated by the fact that the interaction points between the cyber and the physical system are limited. In addition, this allows a tighter integration of safety risk analysis and security risk analysis.

Another challenge is that the typical computation of risk (defined as cost times probability of occurrence) is not a very good metric for security, since the probabilities are unknown, and often hard to estimate. This leads to a very low return-on-investment, which makes the improvement of security a problematic issue. However, although many different attacks can have low probability, the attacker needs to exploit only one of them. Risk assessment should take into account the physical components and the safety risks associated with them.

The participants identified several classes of security consequences:

- Blackout or shutdown of production process
- Damage to the environment/equipment
- Non-optimal operation (higher cost/lower product quality) within production parameters
- Forcing the use of backup processes that are lower quality

Several concrete ideas to improve risk assessment were proposed in the working group.

First, risk analysis should include the cost for the attacker, and identify measures that increase this cost significantly. This especially includes the analysis of potential feedback channels for the attacker; it should be difficult for the attacker to determine the state of the system, so that she cannot immediately determine the success of different phases of the attack. This requires the attacker to build a testbed (as done with Stuxnet).

A second proposal was to reason from the physical system and determine the potential points of attack. This allows integration with safety analysis on the one hand, and development of more resilient physical systems on the other.

Standardization to provide better generic risk assessments and security in general. Current standards (e. g., Modbus) allow custom systems on top of them, and this makes interoperability hard. Interoperability tests do exist, but differing implementations of standards make both security and risk assessment harder and more costly.

One point of disagreement in the discussions was how useful techniques like threat analysis and attack trees are. These approaches are commonly employed for risk analysis in IT-security, but they may not transfer to ICS as well as expected, due to the highly targeted attacks that should be considered. On the other hand, these methodologies allow the use of standardized risk assessments.

5.1.3 Conclusion and Open Challenges

The working group discussion showed that risk analysis for security in ICS requires specific attention beyond employing standardized techniques from IT-security. Several challenges for risk assessment were identified. First, the fact that different stakeholders have different inputs, which leads to widely different risks for different assets, was identified as an open question. Second, the disconnect between risk assessments for safety and for security, respectively for the physical and the cyber component, is an area requiring improvement. Furthermore, it appears that it is especially hard to model risks that do not lead to a conclusive damage, but rather to non-optimal operation. Finally, there was strong criticism expressed regarding the reliance on attack trees, especially considering the fact that there are very few incidents that can be used as a basis for estimating attack probabilities.

5.2 Industry 4.0

Moderation: Nils Aschenbruck (Universität Osnabrück, DE)

Minutes: Alfonso Valdes (University of Illinois – Urbana, US)

License © Creative Commons BY 3.0 Unported license
© Nils Aschenbruck and Alfonso Valdes

This working group addressed the rise of “Industry 4.0”, the associated diversification of industrial control systems (ICSs) and cyber-physical production systems (CPPSs). ICSs are characterized in this context as cyber monitoring and control of processes that interface with the physical world. They are targeted specifically at the industry market, which excludes for example medical devices. Some claim that compared to SCADA, ICSs refer to a more distributed management architecture, although this was controversial. Industry 4.0 refers to a broader and more research-oriented field, although it is still targeted at industry. The term refers to a next generation of factories and farms, which are characterized by highly automated and highly customized networked machines. Customization may even include a end-user specification of products, with checks that the final product meets constraints. Lastly, cyber-physical production systems (CPPSs) are also production- and processing-oriented; some argued they are the same as ICSs. Overall, the exact terminology that should be used is not universally agreed upon and changes over time – therefore, the working group put a focus on use cases that are appropriate for the study of these systems.

5.2.1 New Security Challenges

One of the core issues for each type of system discussed in the working group is the rise of new security challenges by the increased degree of networking introduced by all these ideas. As discussed in the plenary sessions, the increasing connectivity of ICSs in the past has already given rise to a multitude of security issues for current systems. It is important for research to preemptively seek out new security challenges that may arise from future developments.

One important class of challenges relates to the physical consequences that attacks may have for the surroundings of a particular system. These consequences can often be considered as distinct from consequences for the controlled process, such as loss of production and loss of efficiency. A traditional example is the blackout scenario of a smart grid: although the loss of production is a serious problem for the energy provider, the much larger risk of a regional or national loss of the grid plays a big role. Similar issues exist in the different use cases that were discussed in this working group; for example, chemical manufacturing may lead to industrial spills, explosions may occur in many different manufacturing scenarios and pesticides may be released improperly in industrialized agriculture systems. Furthermore, critical infrastructures such as dyke control systems can fail to prevent or even amplify natural disasters such as floods. Traffic accidents are another risk that is associated with many of these use cases. These separate physical consequences are also referred to as societal harm. There is a fundamental distinction between the loss of production, such as manufacturing and agriculture systems, and the loss of control, which is a failure to prevent a disaster.

Privacy is another upcoming challenge for many systems, particularly those that interact closely with end-users. Examples include traffic monitoring systems, smart grids and building automation systems, all of which are typically operated by commercial or sometimes governmental entities. However, research has shown that the sensors in these types of systems can easily be used to deduce significant parts of peoples' lives. On the other hand, the collected data is often essential to the functionality offered by those systems. An open question is who owns this data, and what a fair trade-off is between user privacy and potential profit.

A related but distinct issue is that of confidentiality of data stored in Industry 4.0 systems. In addition to the customization data related to specific end-users, which relates closely to privacy, many ICSs transmit and store a huge amount of data about the monitored process. As the process is often central to the production of a product, this data is of significant value for industrial espionage. The data presents a new way through which company proprietary information may leak into the hands of an adversary. Protecting this information may be a key element in motivating strong security for Industry 4.0.

A final challenge that was discussed is the scalability of existing security mechanisms. Due to the extreme decentralization that is often associated with Industry 4.0, many existing security mechanisms may no longer be sufficient. For example, distributing and regularly exchanging key material and certificates for each and every sensor in a large factory is a difficult process to scale. In the case of industrialized agriculture, compromise of individual sensors or actuators is nearly unavoidable, due to the highly distributed nature of these components, both in space and in sheer number of networked devices. In addition to this challenge, data replication and synchronization will play a role, especially in scenarios where ad-hoc networks are used, as they may not be permanently connected.

To avoid a strong focus on the electrical grid use case, which is a common example, some additional usage scenarios covered by Industry 4.0 were discussed, as their requirements may differ significantly from the electrical grid. The discussed scenarios included future factories

and production systems, environmental monitoring and control, future farms and transport monitoring and control. These scenarios were contributed back to the plenary discussions and were used to avoid a strong focus on the electrical grid use case, whose control loop has very specific properties.

Future factories will feature enhanced customization for customers through increased intelligence in the individual components performing the production process. By performing small changes to the parameters provided to these components, a highly customizable process is created. Important applications of this include customized cars, planes and kitchens, which can be produced more effectively this way. Similarly, multiple companies may order customized components from a single manufacturer. In such a case, there is a significant risk of leaking proprietary information. The potential for collaboration between companies to increase the attack surface of a third company is also a risk that should be considered. A potential solution may be cooperative defense through specifically designed security policies.

5.3 (Real-time) Detection of CPS Attacks

Moderation: Alfonso Valdes (University of Illinois – Urbana, US)

Minutes: Rens van der Heijden (Universität Ulm, DE)

License © Creative Commons BY 3.0 Unported license
© Alfonso Valdes and Rens van der Heijden

5.3.1 Problem Statement

This working group addressed several questions surrounding attacks on cyber-physical systems. The primary proposal for the working group was to address network attacks on industrial control systems, which was limited to real-time network-oriented detection to prevent overlap with the working group on (mainly host-based) forensics.

5.3.2 Discussion Topics

The discussion was started with a follow-up from the results of a proceeding working group on Industry 4.0 and diversity of ICS, to avoid an extensive discussion on which systems we should discuss. From that session, we took several example scenarios: advanced industry, environmental control, future farms and transportation, which served as a guideline for the variety the discussion should address. This turned out to be an important point, due to the fact that many process-centric detection solutions cannot cope with all these scenarios, but rather need to be defined for each use case separately.

A recurring theme in these discussions was identified to be the question of separation between safety and security issues. Purely safety issues should be addressed by control engineers, not by the security community. It was proposed that the security community should avoid involving itself too closely with the specifics of process control, as this leads to several problematic issues. Not only does this lead to a strong dependency on the process, meaning mechanisms need to be customized per process or even per individual factory or production work flow, but the security community specifically needs to avoid building a shadow SCADA system. Finally, effective reaction to security incidents seemed a major open question. Thus, three different general issues were identified at the start of the discussion:

1. combining the constrained environment with standard security management and best practices

2. the availability of semantic knowledge of the process
3. how to effectively and proportionally react to (potential) security incidents

5.3.3 Security Management

The effective deployment of security management has been an active topic in CPS for quite some time. Common problems and example scenarios of failure are often related not to specific attacks on the CPS, but rather relate to poor security management; lack of isolation where appropriate, lack of updates, poor security policies and lack of security awareness. Big examples like Stuxnet are a demonstration of the potential of targeted attacks, but the vast majority of real world compromises are due to old viruses that still affect unpatched systems. This issue strongly relates to the age of the deployed systems, backwards compatibility requirements and the fact that updating often violates guarantees provided by the manufacturer of the system. The working group discussed briefly about why general purpose CPUs/OSs are used, rather than proprietary systems, if generic vulnerabilities are really the cause of the vast majority of incidents. However, it quickly became clear that this was necessary mainly due to cost and tool availability, and it was additionally concluded that proprietary systems could not have offered better security anyway.

5.3.4 Semantic Knowledge

The use of process semantics seems to be an attractive way to further improve detection rates. Because this requires sufficient redundancy in measurements, studying the placement of additional sensors also seems in scope. However, a strong criticism against this is that this is not the domain of security. Rather, safety and process engineers should be responsible for sufficient redundancy in messages and for the detection of safety-relevant events, regardless of whether they come from an attacker or random failures. This view was somewhat controversial, but supported by the fact that incrementally adding process semantics to intrusion detection will lead to an increase in system complexity, and will eventually result in the mirroring of the SCADA system, referred to as a shadow SCADA system. Although some discussion pointed out that mirroring the process monitoring is an approach used by safety and for fault tolerance, the step towards Byzantine fault tolerance was considered far too expensive, due to the fact that $3n + 1$ sessions are required to tolerate n Byzantine faults.

5.3.5 Effective Response

In regular IT security, responses to attacks can often include simply disabling or isolating the affected system, such as a compromised web server, from the network. However, for CPS attacks, this approach is not a reasonable solution, since this response can often lead to significant financial damage. In addition, the attacker can likely exploit such intrusion responses by designing attacks specifically to trigger these responses. In such a case, the response may cause more damage than it prevents

5.3.6 Conclusion and Open Challenges

We identified a fundamental question; is detection of exploits on the physical component, as well as effective placement of additional (physical) sensors, an IT security problem? As raised in the discussion, there are many arguments supporting the use of information about the process to improve detection rates. On the other hand, we need to be weary of the shadow SCADA problem – the security system should not replicate SCADA functionality. In the

end, we agreed that significant disagreement exists on whether process-aware detection is a good idea.

Apart from this result, we also concluded that in reality, many of the research issues we are discussing are not (yet) a problem. Rather, the current challenges faced by the industry are mainly regarding more general IT security challenges, many of which can be found by security consultancy and subsequently solved by standard IT security measures. However, effective deployment of these measures and practices is expensive and challenging due to scaling issues and certification. One potential source of improvement is the certification of processes to change running systems, so that guarantees from manufacturers are preserved while allowing to implement standard security practices to keep systems up to date.

5.4 Cyberforensics

Moderation: Heiko Platzlaff (Siemens – München, DE)

Minutes: Stephan Kleber (Universität Ulm, DE)

License © Creative Commons BY 3.0 Unported license
© Heiko Platzlaff and Stephan Kleber

Proposed discussion topics included:

- Forensic readiness in ICS
- Machine Learning in forensics (ML)
- Forensics of operational data
- Live (monitoring, SIEM) vs. dead (post mortem) forensics where crucial problems of pro-actively making systems forensic ready are non-disturbance of business processes, volatility, influence on operational data
- Differences of forensics between ICS and conventional IT systems
- Safe and secure storage for forensic data is missing
- Is it about discerning host from network?

5.4.1 Abstract

The working group “Cyberforensics” did pursue the question about how forensic analyses in the ICS context can be made more efficient, or even feasible in the first place. One option is to employ methods of machine learning to automate certain steps of the process. This is especially useful at the beginning of an analysis, by filtering out the data to be investigated manually afterwards. Another discussion is the role of SIEM in forensics to correlate incidents, even with data not present in PLCs any more. Based on that is the ranking of the importance of a certain entity for a successful forensic analysis. Finally the central difference in terms of forensics between conventional systems and ICS can be a chance: Context aware analysis makes it feasible to take semantics of operational data into account. Especially PLCs, however, need to be made forensics ready first, for the best of results.

5.4.2 Machine Learning

There are not many approaches known to us about machine learning (ML) in forensics, and conventional strategies are not applicable. However there are enough raw data points to enable ML, but it is learning with few labels resulting in semi-supervised learning strategies.

Starting points for ML could be to learn signatures for software. This resembles an “intrusion detection system (IDS) post mortem” in a manner of speaking, although this point of view is criticized in the IDS community. On the other hand side effects on log files need to be collected. Finally the algorithm should come up with five to six most important events identifying “very likely intrusions” to be reviewed by the forensic analyst. For an informed decision on the likelihood of a false positive for the current issue, the analyst needs to know the details of the reasoning that lead to the conclusion of the ML algorithm. The approach is per se not false positive free but manual review is necessary nevertheless. This allows for higher false positive rates (10 % and above) than are generally allowed for IDS. The main goal here is to reduce the amount of data to be reviewed by filtering. A good basic property for filtering are timestamps of file changes, i. e., how they correlate to genuine updates.

Under the assumption the majority of systems is clean, this majority can be used as a baseline for heuristically determining that and possibly where something went wrong. The alternative approach to infect and replay a specific instance of an previous attack is not difficult in itself, but laborious for which reason it is only rarely performed.

Filtering using persistence, time and signature filters brings down the number of findings to a small amount of files, i. e., for one incident some tens of files. Then typical modes of operation of malware can be a good starting point for further analysis or filtering. It might get necessary to extend the time frame of such filtering gradually. The algorithm is thought to rate changes by their age and descent deeper into the past until an event corresponding to the trigger of the analysis can be identified. Usually there actually is a time-based trigger restricting the time frame of the search.

Currently ten to twelve filters on different features are enough for a useful reduction of data. The relative importance of features may be determined by ranking from ML. There has been done numerous work on classification of malware that resembles the general ranking problem here.

Encryption of executables in general is a problem for ML, whereas PE headers (Windows executable file headers) cannot be encrypted to remain executable. Even most changes typical for obfuscation and packers are detectable through characteristic changes in the header. In general a mismatch of parameters results. For example a low number of imports for a large executable file size is suspicious. False positives can arise when reverse engineering protection mechanisms are in place, that function similarly to a malware packer.

Filter layers that might be evaluated sequentially for more details potentially are: Network, Files, Host, SIEM.

ML is not a silver bullet, but it is able to lead the way by pointing towards important details. Dynamic analysis in this setting might show active parts of an entity. Forensics may for example identify executables in unusual places in memory. But here the important question to answer is, how to find the boundaries for discerning between “unusual” and “usual” in memory analysis.

5.4.3 The Role of SIEM in Forensics

Live analysis of an ongoing attack is easier and needs less effort than static analysis afterwards. Forensic action is generally regarded as reactive and post mortem, in contrast to IDS. But evasion strategies of attackers on the host might cover actions and prevent forensic analysis. Moreover, resource constrained systems prevent a gapless monitoring during operation. Network traces done by a network data collector like a SIEM could be used to fill those gaps, but the general assumption about SIEM is that it will alert live if something is going on. If this holds, no new information may come from this analysis of SIEM traces. Moreover SIEM

is highly dependent on its configuration.

SIEM may provide an inventory of entities, which may be hosts, network components and the like. Each of those can be assigned a priori with a rating of importance for the operation of the whole system. Based on that information the forensics team may be able to find anomalies more targeted.

In this respect it is questionable whether network and host forensics are two disjunct things. Multiple things may have been happening at different places at the same time.

5.4.4 Correlating Time and Events

As discussed earlier an analyst needs help to be able to make informed decisions about possible findings based on the rating of the filtering and ML algorithm. SIEM already is a system based on expert knowledge and rules derived of that. This information about the set of rules that matched for an incident can be given to the analyst.

Based on time-event and time-time correlation, indirect correlations can be revealed. More can be added by considering other features, like user-IDs involved in a file change event. Going beyond that, causal relations of events can be useful to be inferred automatically. But it is unclear whether this even is possible in full extent. Maybe the typical human approach might help to look at: A human would identify and extract one feature at a time to go after in the system, pointing to important places to do actual forensics at. This might then even lead to the analysis and correlation of an event in SIEM or historical network data to backtrack related events.

6 Open Issues

In the final discussion, the seminar led to a number of important conclusions and open research challenges that participants agreed should provide important directions for the future of ICS security research and practice:

- Can Intrusion Detection Systems actually provide better security by becoming “process-aware”, i. e., have detailed information about the process that the ICS controls? While this seems intuitive, others argue that everything done in this direction is simply replicating the control system and provides redundancy but not necessarily better security.
- The interaction between safety and security mechanisms is an important aspect and needs further analysis. While today often treated separately, we think that both areas should work more closely together to work on unified mechanisms.
- ICS, also those beyond Critical Infrastructures, should generally have ‘last-line-of-defense’ monitoring and safety mechanisms that are not connected and not coupled with the potentially attackable ICS. Those mechanisms should provide a ground truth to operators and prevent the system from entering clearly forbidden states.
- Proper reactions to attacks are often very hard to determine for ICS, as a sudden shutdown or disconnection may not be a viable option. ICS security mechanisms like Intrusion Detection and Prevention Systems should therefore be able to provide a flexible reaction to detected security breaches to allow a form of “graceful degradation”. So as in the case of safety mechanisms, ICS should enter more robust and fail-safe states when attacks are detected, perhaps to the detriment of efficiency and output of the controlled process.
- More attention should be paid to user interfaces of security mechanisms to allow operators and security experts appropriate analysis and reaction if attacks cause critical situations.

- Security systems should provide more fine-grained output to allow better forensics and proper reaction to incidents.
- We identified a huge gap between ICS security research in academia and industrial practice. While research targets highly sophisticated attacks and countermeasures, many real-world deployments fail because of lack of even the most simple security best-practices. Closing this gap will require a huge effort that should start with identifying which best-practices have to be applied and which do not fit.
- ICSs also pose big challenges for security management because of the huge scale of some installations, the lack of realistic attacker models that would allow one to find the right level of security, and the economic pressure to build cost-effective security solutions.
- In general, diversity and redundancy are good for ICS security. If a large number of ICSs are from a single vendor and use only one brand of devices, attacks and malware can easily spread and create huge damage. It is therefore not clear yet, whether convergence of ICS to a few vendors and standards (in terms of protocols, operating systems, etc.) will provide more benefits to attackers or to defenders.
- The fact that ICSs are often very long-lived installations and that duration of innovation cycles in ICSs is very different from ICT creates huge problems for maintaining ICS security. Well-defined, certified update processes that are guaranteed for the lifetime of ICSs would significantly support security. However, maintaining own ICS software ecosystems also has economic consequences.
- Separation and isolation (like air gaps, virtualization, sandbox, VPNs) are likely the most effective security mechanisms for ICS. As a corollary, this means that multi-stakeholder ICS like power grids are inherently harder to secure, as they require more interfaces between parties.
- While ICS is a very broad term and encompasses a lot of extremely heterogeneous types of systems, participants were confident that the security challenges to be addressed are often very similar and thus that there can be meaningful progress on ICS security in general without the need to divide the field into further sub-disciplines.

We hope that this list can provide beneficial input to the field and pave the way for future research leading to more secure Industrial Control Systems.

We thank the seminar participants for their active participation and the fruitful discussions. Their contributions formed the basis for the findings presented here.

Participants

- Ali Abbasi
University of Twente, NL
- Magnus Almgren
Chalmers UT – Göteborg, SE
- Nils Aschenbruck
Universität Osnabrück, DE
- Gunnar Björkman
ABB – Mannheim, DE
- Damiano Bolzoni
University of Twente, NL
- Alvaro Cárdenas Mora
University of Texas at Dallas, US
- Marco Caselli
University of Twente, NL
- Jorge R. Cuéllar
Siemens AG – München, DE
- Hervé Debar
Télécom & Management
SudParis – Evry, FR
- Sven Dietrich
City University of New York, US
- Ulrich Flegel
Infineon Technologies –
München, DE
- Dina Hadziosmanovic
TU Delft, NL
- Frank Kargl
Universität Ulm, DE
- Stefan Katzenbeisser
TU Darmstadt, DE
- Richard A. Kemmerer
University of California – Santa
Barbara, US
- Stephan Kleber
Universität Ulm, DE
- Hartmut König
BTU Cottbus, DE
- Marina Krotofil
Hamburg University of
Technology, DE
- Pavel Laskov
Universität Tübingen, DE
- Michael Meier
Universität Bonn, DE
- Simin Nadjm-Tehrani
Linköping University, SE
- Heiko Patzlaff
Siemens – München, DE
- Andreas Paul
BTU Cottbus, DE
- Konrad Rieck
Universität Göttingen, DE
- Rene Rietz
BTU Cottbus, DE
- Robin Sommer
ICSI – Berkeley, US
- Radu State
University of Luxembourg, LU
- Jens Tölle
Fraunhofer FKIE –
Wachtberg, DE
- Alfonso Valdes
University of Illinois – Urbana
Champaign, US
- Rens van der Heijden
Universität Ulm, DE
- Alexander von Gernler
genua Gesellschaft für Netzwerk-
und Unix-Administration mbH –
Kirchheim bei München, DE
- Stephen Wolthusen
Royal Holloway University of
London, GB & Gjøvik University
College, NO
- Emmanuele Zambon
SecurityMatters B. V. –
Enschede, NL



Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities

Edited by

Chris Biemann¹, Gregory R. Crane², Christiane D. Fellbaum³, and Alexander Mehler⁴

1 TU Darmstadt, DE, biem@cs.tu-darmstadt.de

2 Tufts University, US, gregory.crane@Tufts.edu

3 Princeton University, US, fellbaum@princeton.edu

4 Goethe-Universität Frankfurt am Main, DE, mehler@em.uni-frankfurt.de

Abstract

Research in the field of Digital Humanities, also known as Humanities Computing, has seen a steady increase over the past years. Situated at the intersection of computing science and the humanities, present efforts focus on making resources such as texts, images, musical pieces and other semiotic artifacts digitally available, searchable and analysable. To this end, computational tools enabling textual search, visual analytics, data mining, statistics and natural language processing are harnessed to support the humanities researcher. The processing of large data sets with appropriate software opens up novel and fruitful approaches to questions in the traditional humanities. This report summarizes the Dagstuhl seminar 14301 on “Computational Humanities – bridging the gap between Computer Science and Digital Humanities”.

Seminar July 20–25, 2014 – <http://www.dagstuhl.de/14301>

1998 ACM Subject Classification I.2.7 Natural Language Processing, J.5 Arts and Humanities

Keywords and phrases Computer Science, Digital Humanities, Computational Humanities, eHumanities, Big Data, Experimental Methods

Digital Object Identifier 10.4230/DagRep.4.7.80

Edited in cooperation with Marco Buehler

1 Executive Summary

Chris Biemann

Gregory R. Crane

Christiane D. Fellbaum

Alexander Mehler

License © Creative Commons BY 3.0 Unported license

© Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler

Motivation

Research in the field of *Digital Humanities*, also known as *Humanities Computing*, has seen a steady increase over the past years. Situated at the intersection of computing science and the humanities, present efforts focus on building resources such as corpora of texts, images, musical pieces and other semiotic artifacts digitally available, searchable and analyzable. To this end, computational tools enabling textual search, visual analytics, data mining, statistics and natural language processing are harnessed to support the humanities researcher. The processing of large data sets with appropriate software opens up novel and fruitful approaches



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

Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 80–111

Editors: Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler



Dagstuhl Reports

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

to questions in the ‘traditional’ humanities. Thus, the computational paradigm has the potential to transform them. One reason is that this kind of processing opens the way to *new* research questions in the humanities and especially for *different* methodologies for answering them. Further, it allows for analyzing much larger amounts of data in a quantitative and automated fashion – amounts of data that have never been analyzed before in the respective field of research. The question whether such steps ahead in terms of quantification lead also to steps ahead in terms of the quality of research has been at the core of the motivation of the seminar.

Obviously, despite the considerable increase in digital humanities research, a perceived gap between the traditional humanities and computer science still persists. Reasons for this gap are rooted in the current state of both fields: since computer science excels at automating repetitive tasks regarding rather low levels of content processing, it can be difficult for computer scientists to fully appreciate the concerns and research goals of their colleagues in the humanities. For humanities scholars, in turn, it is often hard to imagine what computer technology can and cannot provide, how to interpret automatically generated results, and how to judge the advantages of (even imperfect) automatic processing over manual analyses.

To close this gap, the organizers proposed to boost the rapidly emerging interdisciplinary field of *Computational Humanities* (CH). To this end, they organized a same-named Dagstuhl Seminar that brought together leading researchers in the fields of Digital Humanities and related disciplines. The seminar aimed at solidifying CH as an independent field of research and also at identifying the most promising directions for creating a common understanding of goals and methodologies.

At the core of the organizers’ understanding of CH is the idea that CH is a discipline that should provide an algorithmic foundation as a bridge between computer science and the humanities. As a new discipline, CH is explicitly concerned with research questions from the humanities that can more successfully be solved by means of computing. CH is also concerned with pertinent research questions from computing science focusing on multimedia content, uncertainties of digitisation, language use across long time spans and visual presentation of content and form.

In order to meet this *transdisciplinary* conception of CH, it is necessary to rethink the roles of both computer scientist and humanities scholars. In line with such a rethinking, computer scientists cannot be reduced to software engineers whose task is just to support humanities scholars. On the other hand, humanities scholars cannot be compelled to construe post-hoc explanations for results from automatic data analysis. Rather, a common vision – shared among both groups of scientists – is needed that defines and exemplifies accepted methodologies and measures for assessing the validity of research hypotheses in CH. This vision motivated and formed a common ground for all discussions throughout the seminar.

Goals and Content of the Seminar

In order to elaborate the vision of CH as a bridge between computer science and the humanities, the seminar focused on questions that can be subsumed under four different reference points of problematizing CH:

1. **The Present State: What works, what does not?**

- Review of the success of the last 10 years of the digital humanities: Can we identify commonalities of successful projects? What kinds of results have been obtained? What kinds of results were particularly beneficial for partners in different areas of

research? Can success in one field be transferred to other fields by following the same methodology?

- Review of the challenges of the last 10 years of the digital humanities: What are recurring barriers to efficient cross-disciplinary collaboration? What are the most common unexpected causes of delays in projects? What are common misunderstandings?
 - What is the current role of computer scientists and researchers in the humanities in common projects, and how do these groups envision and define their roles in this interplay?
2. **Computational Challenges in Computational Humanities:**
 - What research questions arise for computational scientists when processing data from the humanities?
 - How can the success of a computer system for humanities data-processing be evaluated to quantify its success?
 - What are the challenges posed by the demands from the humanities? In particular, how can computer scientists convey the notion of uncertainties and processing errors to researchers in the humanities?
 3. **Humanities Challenges in Computational Humanities:**
 - What research questions can be appropriately addressed with computational means?
 - How can we falsify hypotheses with data processing support?
 - What is and is not acceptable methodology when one relies on automatic data processing steps?
 4. **Common Vision: Algorithmic Foundations of Computational Humanities:**
 - Can we agree on generic statements about the expressivity of the range of algorithms that are operative in the digital humanities and related fields of research?
 - Can we distinguish complexity levels of algorithms in the computational humanities that are distinguished by their conditions of application, by their expressiveness or even explanatory power?
 - Which conditions influence the interpretability of the output generated by these algorithms from the point of view of researchers in the humanities?

The Program

In order to work through our set of goals (see Section 1), the seminar decided for a mixture of talks, working groups and plenary discussions. To this end, four Working Groups (WG) have been established whose results are reported in respective sections of this report:

- The Working Group on *Ethics and Big Data* (members: Bettina Berendt, Chris Biemann, Marco Buehler, Geoffrey Rockwell, Joachim Scharloth, Claire Warwick) discussed a very prominent topic with direct relationships to recent debates about ethical and privacy issues on the one hand and the hype about big data as raised by computer science on the other. One emphasis of the WG was on teaching how to process big data, how this research relates to legal and ethical issues, and how to keep on public dialogs in which such issues can be openly discussed – beyond the narrow focus of the academic community. A central orientation of this discussion was to prevent any delegation of such discussions to closed rounds of experts (‘research ethics boards’) which do not support open discussions to a degree seen to be indispensable by the WG. The widespread, fruitful and detail-rich discussion of the WG is reported in more detail in Section 4.1.

- The Working Group on *Interdisciplinary Collaborations – How can computer scientists and humanists collaborate?* (members: Jana Diesner, Christiane Fellbaum, Anette Frank, Gerhard Heyer, Cathleen Kantner, Jonas Kuhn, Andrea Rapp, Szymon Rusinkiewicz, Susan Schreibman, Caroline Sporleder) dealt with opportunities and pitfalls of cooperations among computer scientists and humanities scholars. The WG elaborated a confusion matrix that contrasts commonplaces and challenges from the point of view of both (families of) disciplines. Ideally, scientists meet at the intersection which challenges both groups of scientists – thereby establishing CH potentially as a new discipline. In any event, this analysis also rules out approaches that reduce either side of this cooperation to the provision of services, whether in terms of computing services or in terms of data provisions. More information about the interesting results of this working group are found in Section 4.2.
- The Working Group *Beyond Text* (members: Siegfried Handschuh, Kai-Uwe Kühnberger, Andy Lücking, Maximilian Schich, Ute Schmid, Wolfgang Stille, Manfred Thaller) shed light on approaches that go beyond language in that they primarily deal with non-linguistic information objects as exemplified by artworks or even by everyday gestures. A guiding question of this WG concerned the existence of content-related features of such information objects that can be explored by computational methods. As a matter of fact, corpus building by example of such artifacts is in many cases still out of reach so that computation can hardly access these objects. Seemingly, any success in ‘computerizing’ research methodologies here hinges largely upon human interpretation. Obviously, this is a predestined field of application of human computation with the power of integrating still rather separated disciplines (e. g., musicology, history of art, linguistics etc.). See Section 4.3 for more information about this promising development.
- The Working Group on *Literature, Lexicon, Diachrony* (members: Loretta Auvil, David Bamman, Christopher Brown, Gregory Crane, Kurt Gärtner, Fotis Jannidis, Brian Joseph, Alexander Mehler, David Mimno, David Smith) dealt with the role of information as stored in large-scale lexicons for any process of automatic text processing with a special focus on historical texts. To this end, the WG started from the role of lexica in preprocessing, the indispensability of accounting for time-related variation in modeling lexical knowledge, the necessity to also include syntactic information, and the field of application of automatic text analysis. Special emphasis was on error detection, correction and propagation. The WG has been concerned, for example, with estimating the impact of lemmatization errors on subsequent procedures such as topic modeling. In support of computational historical linguistics, the WG made several proposals on how to extend lexica (by morphological and syntactical knowledge) and how to link these resources with procedures of automatic text processing. See Section 4.4 for more information about the results of this WG.

Part and parcel of the work of these WGs were the plenary sessions in which they had to present their intermediary results in order to start and foster discussions. To this end, the whole seminar came together – enabling inter-group discussions and possibly motivating the change of group membership. Beyond the working groups, the work of the seminar relied on several plenary talks which partly resulted in separate position papers as published in this report:

- In his talk on *Digital and computational humanities*, Gerhard Heyer shed light on the role of computer science in text analysis thereby stressing the notion of exploring knowledge or text mining. He further showed how these methods give access to completely new research questions in order to distinguish between (more resource-related) *Digital Humanities* and (algorithmic) *Computational Humanities*.

- In his talk, Chris Biemann tackled the field of *Machine Learning* methods from the point of view of their application to humanities data. He clarified the boundedness of these methods in terms of what is called understanding in the humanities. From this point of view, he pleaded for a kind of methodological awareness that allows for applying these methods by clearly reflecting their limitations.
- In their talk on *On Covering the Gap between Computation and Humanities*, Alexander Mehler & Andy Lücking distinguished differences that put apart both disciplines. This includes a methodological, a semiotic and an epistemic gap that together result via an interpretation gap into a data gap. In order to overcome these differences, they pleaded for developing what they call hermeneutic technologies.
- In her talk on *Digital Humanities & Digital Scholarly Editions*, Susan Schreibman gave an overview of her work on multimodal, multicodal digital editions that integrate historical, biographical and geographical data. Her talk gave an example of how to pave the way for a people's history in the digital age. To this end, she integrates recent achievements in data mining (most notably network analysis, geospatial modeling, topic modeling and sentiment analysis).
- In his talk on *How can Computer Science and Musicology benefit from each other?*, Meinhard Müller switched the topic of mainly textual artifacts to musical pieces and, thus, to musical artworks. He explained the current possibilities of automatic analysis of musical pieces and demonstrated this by a range of well-known examples of classical music.

This work nicely shows that computational humanities has the goal of covering all kinds of data as currently analyzed and interpreted in the humanities (see also the Working Group *Beyond Text* for such a view).

The seminar additionally included a range of short talks in which participants presented state-of-the-art results of their research: among others, this included talks by Christopher Brown, Anette Frank, Brian Joseph and Szymon Rusinkiewicz. This work nicely provided information about a range of linguistic and multimodal application areas and, therefore, reflected the rich nature and heterogeneity of research objects in the humanities.

A highlight of the seminar was a plenary discussion introduced by two talks given by Gregory Crane and by Manfred Thaller. These talks started and motivated an academic verbal dispute in which, finally, the whole seminar participated in order to outline future challenges of Digital Humanities with impact beyond the border of these disciplines – even onto the society as a whole. Both talks – on *Evolving Computation, New Research Directions and Citizen Science for Ancient Greek and the Humanities* by Gregory Crane (see Section 5.1) and on *The Humanities are about research, first and foremost; their interaction with Computer Science should be too* by Manfred Thaller (see Section 5.2) – opened a broad discussion about the role of humanities among the sciences and their status within the society.

Last, but not least, we should mention two common sessions with a concurrent seminar on Paleography. These sessions, which took place at the beginning and at the end of the seminars, opened an interesting perspective on one particular field that could be counted as a sub-discipline of Computational Humanities. The paleographers met in Dagstuhl for the second time and discussed some of our CH issues previously; it was fruitful to exchange approaches on how to overcome them.

Conclusion

Most of the working groups used their cooperation as a starting point for preparing full papers in which the theme of the group is handled more thoroughly. To this end, the plenary discussed several publication projects including special issues of well-known journals in the field of digital humanities. A further topic concerned follow-up Dagstuhl seminars. The ongoing discussions around the perceived gap between computer science and the humanities and the various proposals from the participants on how to define, bridge or deny this gap made it clear that the seminar addressed a topic that needed discussion and still needs discussion. The talks, panels and working group discussions greatly helped in creating a better mutual understanding and rectifying mutual expectations.

In a nutshell: the participants agreed upon the need to continue the discussion since CH is a young and open discipline.

2 Table of Contents

Executive Summary

Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler 80

Overview of Talks

Digital and computational humanities <i>Gerhard Heyer</i>	87
Design Principles for Transparent Software in Computational Humanities <i>Chris Biemann</i>	88
On Covering the Gap between Computation and Humanities <i>Alexander Mehler and Andy Lücking</i>	91
How can Computer Science and Musicology benefit from each other? <i>Meinard Müller</i>	92

Working Groups

Report of Working Group on Ethics and Big Data <i>Bettina Berendt, Geoffrey Rockwell</i>	94
Report of Working Group on Interdisciplinary Collaborations – How can computer scientists and humanists collaborate? <i>Jana Diesner</i>	96
Report of Working Group <i>Beyond Text</i> <i>Andy Lücking</i>	98
Report of Working Group on Literature, Lexicon, Diachrony <i>Loretta Auvil, David Bamman, Christopher Brown, Gregory Crane, Kurt Gärtner, Fotis Jannidis, Brian Joseph, Alexander Mehler, David Mimno, and David Smith</i> .	99

Panel Discussions

Evolving Computation, New Research Directions and Citizen Science for Ancient Greek and the Humanities <i>Gregory R. Crane</i>	107
The Humanities are about research, first and foremost; their interaction with Computer Science should be too. <i>Manfred Thaller</i>	108

Participants	111
-------------------------------	-----

3 Overview of Talks

3.1 Digital and computational humanities

Gerhard Heyer (Universität Leipzig, DE)

License © Creative Commons BY 3.0 Unported license
© Gerhard Heyer

Joint work of Gerhard Heyer, Volker Boehlke

As manifold as the usages of language are the purposes of text. But when looking at text in the Humanities, it looks to me as a Computer Scientist that we are, broadly speaking, always assuming that the texts we are interested in are encodings of knowledge (of a culture at a time). And this is what makes texts the subject of analysis: By looking at texts (and sometimes also at their context of origin) we intend to decipher the knowledge that they are encoding. Looking at texts from a bird's eye view or taking a close reading perspective has always been the core business of text oriented Humanities. With the advent of Digital Humanities, however, we can scale up this task by using new analysis tools derived from the area of information retrieval and text mining. Thereby all kinds of historically oriented text sciences as well as all sciences that work with historical or present day texts and documents are enabled to ask completely new questions and deal with text in a new manner. In detail, these methods concern, amongst others,

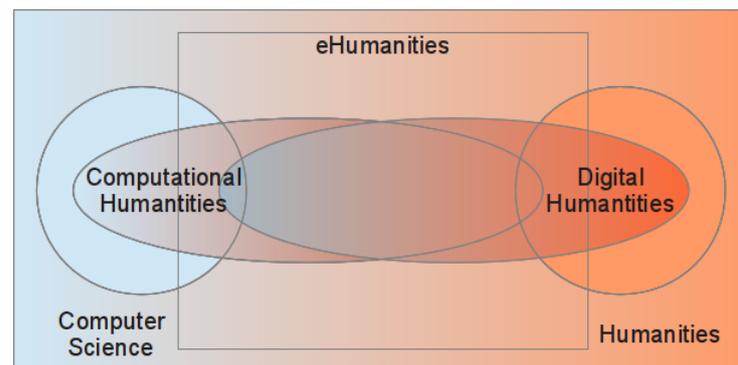
- the qualitative improvement of the digital sources (standardization of spelling and spelling correction, unambiguous identification of authors and sources, marking of quotes and references, temporal classification of texts, etc.);
- the quantity and structure of sources that can be processed at scale (processing of very large amounts of text, structuring by time, place, authors, contents and topics, comments from colleagues and other editions, etc.);
- the kind and quality of the analysis (broad data driven studies, strict bottom-up approach by using text mining tools, integration of community networking approaches, contextualization of data, etc.).

While Computer Science and Humanities so far have acted in their working methodologies more as antipodes rather than focusing on the potential synergies, with the advent of Digital Humanities we enter a new area of interaction between the two disciplines. For the Humanities the use of computer based methods may lead to more efficient research (where possible) and the raising of new questions that without such methods could not have been dealt with. For Computer Science, turning towards the Humanities as an area of application may pose new problems that also lead to rethinking present approaches hitherto favoured by Computer Science and developing new solutions that help to advance Computer Science also in other areas of media oriented applications. But most of these solutions at present are restricted to individual projects and do not allow the scientific community in the Digital Humanities to benefit from advances in other areas of Computer Science like Visual Analytics.

In consequence, I think it is important that we distinguish between two important aspects:

1. the creation, dissemination, and use of digital repositories, and
2. the computer based analysis of digital repositories using advanced computational and algorithmic methods.

While the first has originally been triggered by the Humanities and is commonly known as Digital Humanities, the second implies a dominance of computational aspects and might thus be called Computational Humanities. To distinguish between both aspects has substantial



■ **Figure 1** Positioning of Computational and Digital Humanities in the context of Computer Science and Humanities.

implications on the actual work carried out. Considering the know-how of researchers and their organizational attachment to either Humanities or Computer Science departments, their research can either be more focused on just the creation and use of digital repositories, or on real program development in the Humanities as an area of applied Computer Science.

A practical consequence also in organizational terms of this way of looking at things would be to set up research groups in both scientific communities, Computer Science and Humanities. The degree of mutual understanding of research issues, technical feasibility and scientific relevance of research results will be much higher in the area of overlap between the Computational and Digital Humanities than with any intersection between Computer Science and the Humanities.

3.2 Design Principles for Transparent Software in Computational Humanities

Chris Biemann (TU Darmstadt, DE)

License  Creative Commons BY 3.0 Unported license
© Chris Biemann

Abstract. In this short statement, the importance of transparent software for humanities research is highlighted. Here, three dimensions of transparency are identified: First, software should be freely available so that results are reproducible. Second, software should be easy to use and hide complex underlying algorithmics from the user. Third, to avoid a black box situation where the software's decisions are opaque to the user, the reasons for any of the automatically produced statements should be traceable back to the data they originated from. After elaborating on these principles in more detail, they are exemplified with a basic distant reading application.

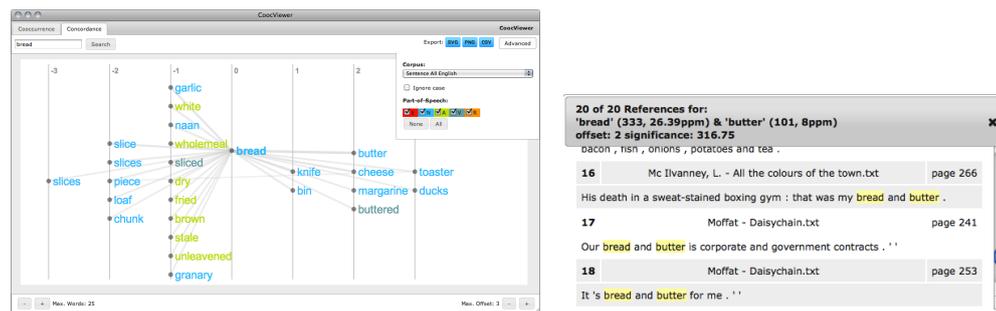
Introduction. The newly emerging field of Computational Humanities (CH) is situated at the interface between humanities research and computer science. Research questions in CH are concerned with aspects of both fields: in Digital Humanities (DH) research, computational aspects either not considered relevant or are merely assigned a subordinated role, while in computer science, research on computational methods and algorithmic approaches is rather detached from their application domain – e.g. the field of Machine Learning produces

methods that learn from data, no matter what kind of data it is. In contrast to this, CH considers humanist's questions and computational challenges both as first-class citizens, and focuses on their interplay. Whereas in both Computational and Digital Humanities, software solutions are needed that support the humanist – typically in accessing electronically available data in her respective field of study – CH research is also concerned with further automatizing the analysis using novel algorithmic approaches. As opposed to generic computer science approaches, however, algorithms in CH software are additionally required to be comprehensible by human(ist)s, in order to be open for scrutiny to allow for a depth of analysis that is satisfactory for the humanities. With respect to these prerequisites, a number of requirements on the software can be deduced. These will be subject of the following section, which discusses three dimensions of transparency that CH software should have in order to be a suitable tool for CH research. On a related topic, but written from the perspective of Computational Linguistics, see Pedersen (2008).

Transparency of Software for Computational Humanities. The term ‘transparency’ can be defined in organizational contexts as ‘the perceived quality of intentionally shared information from a sender’ (Schnackenberg & Tomlinson, 2014) and implies openness, communication and accountability. In this section, these facets of transparency are elaborated on and put forward as desired properties of software used in Computational Humanities research.

Open Source for Reproducibility. Whether hypotheses are merely empirically verified on data that has been mined by computational approaches, or hypotheses are generated from empirical observations in the first place: research in CH inherently includes empirical aspects, and rational deduction is complemented by a certain amount of experimentation. As in the experimental sciences, such as e.g. Physics, empirical investigations in CH must be reproducible to adhere to scientific standards. Just as it is considered bad science in the field of computational linguistics to rely on commercial search engines for data acquisition and statistics (Kilgarriff, 2006) because their inner workings are secret and they change over time, the CH researcher should not rely on commercial software with closed sources for the same reason. Rather, software in CH and other research contexts should be available open source in versioned public repositories, and the version of the software should be included in the description of the experimental setup. In this way, subsequent research is able to reproduce prior experiments of others and the inner workings of the software are fully transparent, at least for those that can understand computer programs. A further advantage of open source software over proprietary software, especially when distributed under a lenient license, is the possibility for subsequent research to combine several existing software into more advanced and more complex software without having to re-implement already existing methods.

Intuitive Interfaces and Hiding Complexity. Just as in communication between humans, communication, i.e. human-computer interaction, happens when a CH researcher uses CH software. And just as successful fact-oriented communication between humans just provides enough detail to communicate the intended amount of information, supportive software should be intuitive to operate and hide unnecessary complex aspects from the user. For this, design principles of graphical user interfaces should be adhered to, and e.g. developed according to the visual analytics process (Keim et al., 2010). Abstracting from complexity, however should not be confused with obfuscation – while it is necessary for the acceptance of the software and its methods that algorithmic results are easy to obtain without necessarily understanding the algorithmic details, it is still crucial that the implementation of such details are transparent (cf. Section 2.1) and the algorithmic decisions are backed up by access to the data that leads to these decisions (cf. Section 2.3). Only in this way, the CH researcher



■ **Figure 2** CoocViewer software, showing significant concordances for “bread” and source text information for “bread and butter”.

can build trust in her algorithmic methodology and develop an intuition about its utility and potential. A result of a successful CH research is always twofold: an algorithmic method and/or a mode of its application that allows to easily analyze data from the humanities, and a result in humanities research obtained with the help of such method.

Accountability and Provenance. The most precise automatic result will still be subject to doubts and disbelief by human experts, as long as no explanation is provided how the automatic method arrived at such result. As mentioned in the previous section, in order for a method to be trusted, it needs to provide the possibility to drill down into the details of its decision-making process, to be fully accountable and to provide a fully transparent reason why the method arrived at a particular result, which is in software development known as data provenance (cf. W3C.org, 2005; Simmhan et al., 2005). In the context of CH, data provenance means not only to store and use algorithmic derivations of the input data (such as e.g. the number of times a certain term appears in texts of a certain time span), but also the sources from which these derivations were derived from (i.e., pointers to the positions in the documents where the term appeared) and a way to access them via the user interface. Data provenance enables the researcher to judge the software’s decisions and to accept or discard algorithmically found evidence.

CoocViewer – a Distant Reading Tool. In this section, we discuss CoocViewer (Rauscher et al., 2013), a simple tool for distant reading, along the three facets of transparency as outlined above. CoocViewer is an Open Source tool that allows browsing of statistically extracted networks of terms (cf. Quasthoff et al., 2006) extracted from corpora in the format of significant concordances. Figure 2 shows significant concordances for the term ‘bread’. The complexity of the computation of such concordances and details of the concordance are abstracted; the user only notices the most significantly co-occurring terms, for example ‘butter’ located two positions to the right of ‘bread’. To investigate this connection, the user can click on the link and drill down into all 20 references that lead to the link, as shown on the right side of the figure: CoocViewer provides full data provenance by showing – on demand – detailed information about single word frequencies and the references, including document titles and page numbers.

While not being a very complex example, CoocViewer adheres to the three design principles of transparency for CH software. Additionally, it enables the import and export of data in various formats for improved usability. During its development, several measures of significance, which determine the related terms shown as most significant concordances, have been examined to investigate computational aspects of distant reading. The tool was productively used in quantitative literary analysis of crime novels, see (Rauscher, 2014).

Conclusion. This short statement laid out design principles for the transparency of software for the computational humanities. Three important facets of transparency were identified that are desirable for software in the field of Computational Humanities: open source codebases for reproducibility, intuitive interfaces for effective communication between user and software, and data provenance for accountability and to build trust in algorithmic methods. These facets were exemplified on CoocViewer, a distant reading tool that adheres to these principles. Creating software to answer research questions in humanities research and computational research alike is one of the main aspects of the field of Computational Humanities. Adhering to the design principles of transparency, as discussed in this statement, enables a firm basis for reproducible research, the exchange of techniques and components, and the credibility of results through data provenance. Thus, not only the source data should be available freely to other researchers, but also the software that allows us to produce scientific results in the field of computational humanities.

References

- 1 Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. (Eds.). *Mastering The Information Age – Solving Problems with Visual Analytics*. Eurographics Association, 2010
- 2 Kilgarrieff, A. *Googleology is Bad Science*. *Computational Linguistics* 33(1):147–151. 2006
- 3 Quasthoff, U., Richter, M. and Biemann, C. *Corpus Portal for Search in Monolingual Corpora*. Proceedings of LREC-06, Genoa, Italy, 2006
- 4 Rauscher, J., Swiezinski, L., Riedl, M., Biemann, C. *Exploring Cities in Crime: Significant Concordance and Co-occurrence in Quantitative Literary Analysis*. Proceedings of the Computational Linguistics for Literature Workshop at NAACL-HLT 2013, Atlanta, GA, USA. 2013
- 5 Schnackenberg, A., Tomlinson, E. *Organizational Transparency: A New Perspective on Managing Trust in Organization-Stakeholder Relationships*. *Journal of Management* DOI: 10.1177/0149206314525202. 2014
- 6 Simmhan, Y.L., Plale, B., Gannon, D. *A Survey of Data Provenance in e-Science*. *ACM SIGMOD Record*, 34(3):31–36, DOI: 10.1145/1084805.1084812. W3C (2005): http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

3.3 On Covering the Gap between Computation and Humanities

Alexander Mehler and Andy Lücking (Goethe-Universität Frankfurt am Main, DE)

License © Creative Commons BY 3.0 Unported license
© Alexander Mehler and Andy Lücking

Since digital or computational humanities (CH) has started its triumph in the humanities' research landscape, it is advisable to have a closer look at its methodological and epistemological range. To this end, we look at CH from the point of view of preprocessing, machine learning, and the general philosophy of science and experimental methodology. From this perspectives, a number of gaps between CH on the one hand and classical humanities on the other hand can be identified. These gaps open up when considering: (i) the status of preprocessing in CH, its logical work-flow and the evaluation of its results compared to the needs and terminological munition of the humanities. Most importantly, corpus preprocessing often comes *before* hypothesis formation and respective model selection has been carried out, turning the logically as well as methodologically required workflow upside down. (ii) The predominant role of functional explanations in CH applications vs. the predominant role of intentional explanations with regard to the humanities. While so far computational

processes can at most be functionally evaluated, hypotheses made in the humanities are usually embedded within contexts of justification that draw on some intentional statement. (iii) The possibilities of falsifying CH hypotheses and hypotheses in the humanities. Given the different typical patterns of explanations (see (ii) above), the results of computations and of the humanities cannot be put to falsification as known from the powerful methodology from the natural, experimental sciences. This leaves open questions about the validity of these results. (iv) The use of big data in CH vs. the use of deep data in the humanities. Analyses in the humanities usually involve the interpretation and rational reconstruction of their objects. This hermeneutic procedure goes beyond mere preprocessing and parsing of those objects, as is typically within reach of CH applications. When gathering interpreted and preprocessed data into corpora (which is done only seldom in the humanities, though), both approaches result in different kinds of resources which may be only of marginal benefit for the respectively other party. (vi) The lack of experimental methods in both CH and the humanities. In order to implement a notion of falsification in CH, one needs to think of CH-specific experimental settings which give rise to test procedures in the first place.

Based on these assessments, we argue that there are at least five interrelated gaps between computation and humanities, namely

1. an *epistemological gap* regarding the kind of evaluation mainly addressed by computational models in contrast to the kind of explanations addressed in the humanities;
2. a *data-related gap* regarding the build-up of ever growing text corpora in computer science in contrast to the need of controlled as well as deeply annotated data in the humanities;
3. a *semiotic gap* regarding signs as strings in the CH in contrast to rich sign-theoretical notions employed in the humanities;
4. a *methodological gap* with respect to understanding the functioning of methods of computer science by humanities scholars; and
5. an *interpretation gap* regarding the foundation of statistical findings in terms of the theoretical terms of the humanities involved.

Having diagnosed these gaps we proceed by delineating two steps that could narrow (some of) these gaps: firstly, the understanding of CH technologies should be fostered by implementing them as part of a curriculum. Secondly, we should think of hybrid algorithmic methods, i. e. methods that at crucial branching points involve humanist expertise from the outset and in this way may pave the way towards “hermeneutic technologies” as a special kind of human-based evolutionary computing.

3.4 How can Computer Science and Musicology benefit from each other?

Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

License © Creative Commons BY 3.0 Unported license
© Meinard Müller

Joint work of Müller, Meinard;

Main reference M. Müller, V. Konz, M. Clausen, S. Ewert, C. Fremerey, “A Multimodal Way of Experiencing and Exploring Music,” *Interdisciplinary Science Reviews (ISR)*, 35(2):138–153, 2010.

URL <http://dx.doi.org/10.1179/030801810X12723585301110>

Significant digitization efforts have resulted in large music collections, which comprise music-related documents of various types and formats including text, symbolic data, audio, image, and video. For example, in the case of an opera there typically exist digitized versions of the

libretto, different editions of the musical score, as well as a large number of performances given as audio and video recordings. In the field of music information retrieval (MIR) great efforts are directed towards the development of technologies that allow users to access and explore music in all its different facets. For example, during playback of some CD recording, a digital music player may present the corresponding musical score while highlighting the current playback position within the score. On demand, additional information about melodic and harmonic progression or rhythm and tempo is automatically presented to the listener. A suitable user interface displays the musical structure of the current piece of music and allows the user to directly jump to any key part within the recording without tedious fast-forwarding and rewinding. Furthermore, the listener is equipped with a Google-like search engine that enables him to explore the entire music collection in various ways: the user creates a query by specifying a certain note constellation, some harmonic progression, or rhythmic patterns, by whistling a melody, or simply by selecting a short passage from a CD recording; the system then provides the user with a ranked list of available music excerpts from the collection that are musically related to the query.

In the Dagstuhl seminar, I have provided an overview of a number of current research problems in the field of music information retrieval and indicated possible solutions. One goal within the Computational Humanities is to gain a better understanding to which extent computer-based methods may help music-lovers and researchers to better access and explore music in all its different facets thus enhancing human involvement with music and deepening music understanding. How may automated methods support the work of a musicologist beyond the development of tools for mere data digitization, restoration, management and access? Are data-driven approaches that can access large amounts of music data useful for musicological research? Vice versa, what can computer scientists learn from historical musicology? How can one improve existing techniques by incorporating knowledge from music experts? How do such expert-based approaches scale to other scenarios and unknown datasets?

References

- 1 Meinard Müller, Michael Clausen, Verena Konz, Sebastian Ewert, Christian Fremerey. *A Multimodal Way of Experiencing and Exploring Music*. *Interdisciplinary Science Reviews (ISR)*, 35(2):138–153, 2010.
- 2 David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, Meinard Müller. *A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction*. *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, 12(2-3):53–71, 2012.
- 3 Meinard Müller, Thomas Prätzlich, Benjamin Bohl, Joachim Veit. *Freischütz Digital: A multimodal scenario for informed music processing*. In *Proceedings of the 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, 2013.
- 4 Verena Konz, Meinard Müller, Rainer Kleinertz. *A Cross-Version Chord Labelling Approach for Exploring Harmonic Structures – A Case Study on Beethoven’s Appassionata*. *Journal of New Music Research*:1–17, 2013.

4 Working Groups

4.1 Report of Working Group on Ethics and Big Data

Bettina Berendt (KU Leuven, BEL), Geoffrey Rockwell (University of Alberta, CAN)

License © Creative Commons BY 3.0 Unported license

© Bettina Berendt, Geoffrey Rockwell

Joint work of Bettina Berendt, Chris Biemann, Marco Büchler, Geoffrey Rockwell, Joachim Scharloth, Claire Warwick

The following is the report of the Working Group on Ethics and Big Data (EBD) at the Dagstuhl seminar on Computer Science and Digital Humanities¹. This working group was formed to discuss ethical and privacy issues around big data following the Snowden revelations. Some of the questions we asked included:

- What are the ethical and privacy issues raised by big data methods?
- What are our responsibilities as researchers and educators working around big data?

We came to the conclusion that, whatever position one might take on the ethics of big data, we have responsibility to expose our students to the lively discussion around the issue. This led to a more focused question:

- How can we teach the ethics of big data?

During the course of our deliberations we did the following:

- We experimented with a close reading of the CSEC slides.² The idea was to use slides leaked by Snowden to both a) explore EBD across disciplinary boundaries and b) to experiment with a way of teaching EBD through current materials. Such close reading of primary source documents about big data and surveillance can bring CH and DH folk together. We need the CH folk to read the software represented and the DH folk to read the documents as rhetorical documents. There is an interesting opportunity also for joint research at this intersection.
- We discussed the literature and archives that need to be explored in this area. (See the Appendix below for some of the archives identified). We agreed to share resources. Rockwell has, for example, create a preliminary reading list to be built on.³
- We agreed to share pedagogical materials. Berendt has shared her materials and other plan to as they experiment with teaching EBD.⁴
- We discussed the development of an infographic that makes the case for the importance of ethics in big data.
- We agreed to develop a web site with resources on this subject. Büchler has set up the basic infrastructure for this and we will begin to populate it as we experiment with teaching EBD.
- We agreed to write a short (5000 word) opinion piece for the “Discussions” column of KI – Künstliche Intelligenz (<http://www.springer.com/computer/ai/journal/13218>). We outlined an argument we were all comfortable with as a way of developing a common ethic. (See Appendix A: Discussion Outline).

¹ See <http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=14301>

² See http://www.scribd.com/fullscreen/188094600?access_key=key-2dvzkv8d3gnowt96adba&allow_share=true&view_mode=scroll

³ See <http://philosophi.ca/pmwiki.php/Main/BigDataEthicsReadings>

⁴ See <http://people.cs.kuleuven.be/~bettina.berendt/teaching/2014-15-1stsemester/kaw/index1.htm> – these materials are used in a course described here <http://people.cs.kuleuven.be/~bettina.berendt/teaching/2014-15-1stsemester/kaw/>

Conclusion: It turns out that the issues are compelling and the reading of original documents like the CSEC slides to understand what the NSA (and others) are doing is one way into a shared discussion about EBD. Some of our conclusions were that:

- There are good ways to get people engaging with the issues – both the issues of ethics and the issues of what be really done get raised.
- We can imagine how we can turn forensic/diplomatic readings into problems for students and a site for interdisciplinary research.
- What the CSEC slides show is a process not unlike what we do ourselves (or want to do). This raises the issue of what the difference is between academic work and SIGINT (Signals Intelligence)? What makes one use of big data methods ethical or not?
- If the big data processes revealed by the Snowden leaks show good (or at least interesting) examples of big data interpretation (or analysis) then can we learn from them? Would it be ethical to copy the tools or processes revealed?
- Ultimately we have to ask how surveillance is different from research or forms of care for the other? Both are a form of knowing another – how is the other different and how is the knowing different?

Appendix A: Discussion Outline

- 1 Introduction (framing it in terms of the current discussion)
 - 1.1 How should we do big/data/science in light of the Snowden revelations?
 - 1.2 Background to Snowden revelations
- 2 What do academics have to offer? What is the role of the researcher now?
 - 2.1 We teach big data,
 - 2.2 We are researchers developing new methods and tools,
 - 2.3 We provide data,
 - 2.4 We are citizens, and
 - 2.5 We can act as mediators
- 3 The standard position on the ethics of big data is that it is not my business – that big data is just a tool/technique
 - 3.1 How do researchers talk about their developments?
 - 3.2 What can we learn from philosophy of technology?
 - 3.3 Mining is about discriminating – one cannot avoid legal and ethical issues
- 4 We should beware inventing the other – we need to ask about our activities in the academy too. How is our research a form of surveillance.
 - 5 Therefore we see the need for public dialogue rather than ethical decision trees that absolve people of the need to think about what they are doing
 - 5.1 Most people think research ethics boards are the solution – but this delegates
 - 5.2 We also need thick description – story telling
 - 5.3 Finally, we need to teach people across humanities/data sciences

Appendix B: Archives and Literature

- ACLU: <https://www.aclu.org/nsa-documents-search>
- Der Spiegel: <http://www.spiegel.de/international/the-germany-file-of-edward-snowden-documents-available-for-download-a-975917.html>
- Nymrod: <http://www.spiegel.de/media/media-34098.pdf>
- Cryptome: <http://www.cryptome.org>
- LeakSource: <http://leaksource.info/category/nsa-files/>

■ **Table 1** Collaboration Scenarios.

	Trivial or old Computing	Challenging or new Computing
Trivial or old Humanities	A) Routine work; Might still be worthwhile pursuing, e. g. for digitizing data or generating data/material for a project	B1) Humanities as a service, B2) Informed reuse of humanities knowledge, material, data (example from comp. linguistics: expertise for data annotation/ markup schemas)
Challenging or new Humanities	C1) Computing as a service, C2) Informed reuse of computing knowledge, data, skills (Example from comp. linguistics: building automatic annotation solutions)	D) Sweet spot of collaboration; Ideal situation: advancing science/ state of the art in both camps

4.2 Report of Working Group on Interdisciplinary Collaborations – How can computer scientists and humanists collaborate?

Jana Diesner (*University of Illinois at Urbana Champaign, US*)

License © Creative Commons BY 3.0 Unported license
© Jana Diesner

Joint work of Jana Diesner, Chistiane Fellbaum, Anette Frank, Gerhard Heyer, Cathleen Kantner, Jonas Kuhn, Andrea Rapp, Szymon Rusinkiewicz, Susan Schreibman, Caroline Sporleder

Our group explored obstacles, solutions and different types of benefits for the collaboration between humanities scholars and computing scholars. We formalized common pitfalls as well as opportunity spaces into a novel framework or model as shown in Table 1. We are currently working on turning this framework and the outcome of our discussion into a paper.

Ultimately, the ideal DH project will land in the lower right corner, where it advances science in the humanities and computing. The upper left corner bears no innovation for either domain, but might be a necessary precondition – e. g. data digitization work – for enabling some subsequent projects. The other two cells will entail innovation majorly for either the humanities or computing; with the other discipline serving as a utility or repository of data or methods. We believe that is essential for a successful project to identify where in this grid it belongs. Our planned paper can then help to identify common challenges and opportunities.

For each cell in this table, we have discussed preconditions, insights from both perspectives, pitfalls aligned with possible solutions and best practices. We believe that such an overview can help scholars to be more systematic and comprehensive in addressing the following problems:

- Identification and definition of the objectives and advantages for everyone involved in a CH/DH project? Possibly different ones for the computing versus the humanities people. A project might not advance both disciplines.
- Norms and standards, e. g. for publishing, co-publishing, performance, data gold standards

- Evaluation
- Barriers to collaboration: fears (traditional humanists: becoming obsolete, computing people: results of too low accuracy to get published, both: requirements associated with complexity of interdisciplinary projects including learning)
- Intellectual property (data, tools)

We have translated this model into implications for the actual process of starting and working through a collaborative project. We have concluded that prior to a project, extensive communication and discussions are needed to clarify on a couple of critical points. These points are outlined below and will be further detailed in our paper:

Identification of which cell a project falls into

- Is everybody ok with that?
- Collect arguments as to why either side could not carry out the work alone

Funding

- Hard for data collection, digitization (cell A)
- Easier for computing people, discuss role of humanities scholars in that case to ensure balanced responsibilities (cells C, D)

Expectations

- Computing "burns methods", Humanists "burn data". This requires a discussion on the value of methods and data, and expected standards for both.

Success criteria for each camp

- Methodological quantitative questions (focus in computing) versus substantive questions (focus in humanities)
- Data: amount, quality
- Performance: What matters? Speed? Accuracy? Theory? Understanding?
- Publishing

Mutual learning

- What amount of learning about knowledge and/ or skills from other camp is a) needed and b) expected?
- What learning resources are available? Add time for training into the grant application? Record training material as reusable resource?
- "Celebrate the gap"? Under what conditions does not each side necessarily need to be intimately familiar with the other

Team composition

- Mediators needed?
- Student research opportunities?

Standards

- Data collection
- Data analysis
- Evaluation
- Level of formalization, generalizability
- Publishing

We will bring the model shown in Table 1 together with these criteria for each cell in the matrix and align them with implications for the steps needed in every research process. Our team entails members from the humanities, computing and both, which we believe is essential for fleshing out these pitfalls and remedies.

4.3 Report of Working Group *Beyond Text*

Andy Lücking (Johann Wolfgang Goethe University Frankfurt am Main, DE)

License © Creative Commons BY 3.0 Unported license
© Andy Lücking

Joint work of Siegfried Handschuh, Kai-Uwe Kühnberger, Andy Lücking, Maximilian Schich, Ute Schmid, Wolfgang Stille, Manfred Thaller

The Working Group *Beyond Text* deals with any kind of media except text (i. e. written language). Accordingly, the group started by enumerating kinds of media as objects for digital humanities (DH). Due to the personal constitution of the group, the prime examples discussed are artworks (primarily paintings) and communicative everyday gestures. The example of paintings leads directly to a huge challenge for the feature-oriented focus of digital, corpus-based methods prevalent in DH: paintings exhibit properties bound up with their *expressiveness* that cannot straightforwardly reduced to (sets of) material features of the paintings – if they can be reduced at all. In particular, aesthetic judgments, for example, draw on normative backgrounds that are not part of the painting proper. As a consequence, such properties are out of reach for computational methods that only have (a digital representation of) the painting in question at their disposal. Such higher-order aspects of images, therefore, still rely on human interpretation, probably made explicit in annotation. Thus, respective work in DH seems to involve a hermeneutic dimension that so far is out of reach of computational automatization. This line of thinking, therefore, pinpoints a gap between humanities and DH and shed some light on a division of labor.

This result leads to curricular issues: what kinds of knowledge and which skills does a DH researcher need to have? Obviously, a genuine DH researcher optimally can decide which part of preprocessing or analysis can be done automatically and which part requires human interpretation. In order to make such a decision, the DH researcher needs to have a basic understanding of DH technology on the one hand, and of the hermeneutic methods in the humanities' discipline in question on the other hand. At this point, a self-evident connection to groups discussing curricular issues emerges.

A particular feature of paintings (though clearly not an exclusive one) is *vagueness*. Accordingly, the group discussed vagueness as a sample topic for DH dealing with media beyond text. Vagueness in paintings comes in a variety of manifestations: the colors of a painting give rise to a gradation known very well from categorization and prototype theory. The painting technique itself (e. g., *sfumato*) may result in a “visual vagueness” due to blurring the depicted scene and thereby preventing a clear recognition. Some features of the text may simply be unknown or uncertain like the name of the painter or the year of painting. Furthermore, paintings often draw on ambiguities of different kinds, ranging from flip-flop images over superimposed encodings to iconographic stylizations on top of figurative painting. A precondition for DH therefore is to distinguish different kinds of vagueness. According to the above-given list, at least the following phenomena have to be distinguished: *epistemic vagueness*, *visual vagueness*, *fuzziness*, *ambiguity*, and *interpretational vagueness*. Whether or not all these phenomena are subsumed under vagueness or a divergent terminological rendering is preferred, DH tools and techniques have to deal with them. This pertains to information storage (databases) as well as to computational modeling (e. g. fuzzy logic).

A special problem in this context is due to logical inconsistencies. Such inconsistencies can be the result of merged perspectives in paintings (think of the famous paintings of Escher) or of conflicting descriptions in texts (for instance, if the protagonist is sometimes described to be a left-hander, other times to be a right-hander). Problems of fictional speech acts and statements in fictional theory aside, a useful DH application has to provide even conflicting

information. Of course, contradictory details can simply be gathered in, say, a database. But this would come at a high prize: the application of inference engines would be blocked. The group discussed some application scenarios and possible technical solutions, though a realizable joint project had to be postponed to further collaboration.

It has to be emphasized that this summary is highly streamlined in the sense that it neither reflects nor exhausts the thematic and rhematic dynamics of discussions. Although only few talking threads converged into a viable proposal, the involvement of discussions shows that there is a great need for exchange of researchers from different backgrounds working in roughly the not yet delineated field of DH.

4.4 Report of Working Group on Literature, Lexicon, Diachrony

Loretta Auvil (Illinois Informatics Institute, Urbana, IL, USA), David Bamman (Carnegie Mellon University, Pittsburgh, PA, USA), Christopher Brown (The Ohio State University, Columbus, OH, USA), Gregory Crane (University of Leipzig, DE, and Tufts University, Medford, MA, USA), Kurt Gärtner (University of Trier, DE), Fotis Jannidis (University of Würzburg, DE), Brian Joseph (The Ohio State University, Columbus, OH, USA), Alexander Mehler (Goethe University Frankfurt, DE), David Mimno (Cornell University, Ithaca, NY, USA), David Smith (Northeastern University, Boston, MA, USA)

License © Creative Commons BY 3.0 Unported license

© Loretta Auvil, David Bamman, Christopher Brown, Gregory Crane, Kurt Gärtner, Fotis Jannidis, Brian Joseph, Alexander Mehler, David Mimno, and David Smith

Joint work of Loretta Auvil, David Bamman, Christopher Brown, Gregory Crane, Kurt Gärtner, Fotis Jannidis, Brian Joseph, Alexander Mehler, David Mimno, David Smith

4.4.1 Introduction

The Working Group on *Literature, Lexicon, Diachrony* identified three key issues or themes that pertain to the computational study of structured linguistic resources (prototypically, the lexicon) and unstructured text. These themes are the following:

- characterizing the nature of the information that has been captured in existing lexica written for human use and the possibilities for rendering these linguistic resources useful for automatic processing;
- exploring the possibilities of creating and augmenting linguistic resources by analyzing texts, and in particular in capturing diachronic variation; and
- analyzing, classifying, and mitigating errors introduced at each stage of processing, from optical character recognition and human annotation, to the construction of word frequency distributions and topic models, to part-of-speech (POS) tagging, lemmatization, parsing, and narrative analysis.

Schematically (as depicted in Table 2), these themes fit within a typology of complementary human and machine annotations. In what follows, we elaborate on each of these themes and develop within each various related sub-issues, some of which overlap with one another or serve as a bridge linking one theme with another.

4.4.2 The Nature of the Lexicon

The value of digitized lexica is well established: even elementary steps of text processing like OCR correction gain a great deal from access to lexica – not to speak of more challenging

■ **Table 2** Stages of lexicon formation contrasted with automatic automatic processing and human annotation.

Stage	Human	Automated
Text creation	Double-keying	OCR
Combining variant forms	Morphology, lemmatization	String-edit clustering, morphological classification, named-entity recognition
Lexical disambiguation	Examples of textual citations, usage	PoS-tagging, contextual clustering
Sense disambiguation	Query expansion from existing definitions, organizing examples into categories	Latent semantic and topic analysis, contextual clustering
Relationships: phrases, synonyms, antonyms, frames, names	Examples of connections between documents	Collocate detection, parsing, lexical patterns (e. g. <i>not just X but Y</i>)

tasks like textual entailment or discourse parsing. Our discussion began by asking what a dictionary is and what purpose it serves. More specifically, we asked whether it is a repository of information, an authoritative statement that users can turn to for answers, a snapshot of a language at a particular point in time, or just what (for a comprehensive international survey of lexica see Hausmann et al. 1989).

For each stage of lexicon creation, there are both manual and automatic methods. We argue that modern workflows should incorporate both types of analysis. Table 2 shows correspondences between methods at each stage.

- **On the value of dictionaries:** There are various types of lexicon/dictionary serving different functions. For literary and linguistic research, lexica/dictionaries on historical principles are essential aids for the diachronic study of texts from the first records of a language up to its present-day varieties. Information technologies can contribute enormously to enhance the uses of existing dictionaries in various ways, thus satisfying the requirements of linguists and philologists studying texts (textual data), words and their histories. (Retro-)digitized lexica/dictionaries play a key role in transforming lexicographical resources from book form with alphabetic macro structures into more efficient means of locating reliable, accurate and comprehensive information; the user is no longer restricted to entries in alphabetical order, but can perform complex searches and exploit all the riches of information stored in a lexicon. The Perseus project⁵ (see Crane 1996, also Lidell & Scott 1996) is one example of this.

In the field of the vernacular languages, the scholar of *Middle High German* (MHG) in pre-electronic times had to use at least four dictionaries for this language period (ca. 1050 up to ca. 1350). These dictionaries have been digitized and all the essential information positions have been encoded carefully in order to allow complex searches related to lemma and word formation, word class, languages of loanwords, diachronic and diatopic features and document types of sources. The digitized dictionaries have been interlinked, so that an entry can be searched in all four lexica displayed synoptically

⁵ <http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057>

on the screen (see: <http://mwv.uni-trier.de>). In the off-line version the search can be restricted to specific sources, e. g. the Arthurian novels, the writing of the mystics etc., or to a single text e. g. the *Parzival* by Wolfram von Eschenbach (see Fournier 2001). Furthermore, the existing MHG dictionaries are interlinked with the new MHG dictionary (*Mittelhochdeutsches Wörterbuch*) which is being published since 2006 in book form and concurrently on the Internet (<http://www.mhdwb-online.de>); for more information about the electronic text archive, lemmatization procedures etc. see Gärtner (2008). The interlinking techniques via normalized lemmata allow for the creation of dictionary nets for a certain period of a language. The period-related subnets can be interlinked with other historical dictionaries of a certain language, e. g. the *Deutsches Wörterbuch* (DWB, 2nd edition DWB²) by the brothers Grimm (<http://dwb.uni-trier.de>). The interlinking can be achieved in various ways, e. g. via period-specific lemma forms in the head of an entry or even by semantic features (see: <http://woerterbuchnetz.de>). An even more global net of dictionaries could comprise dictionaries through more subnets e. g. for the Germanic languages: Gothic, Old English, Old Saxon, Old High German and Old Norse. Interrelations of language stages, linguistic borrowings etc. can be studied in new and more reliable ways, if linguists and philologists are willing to look over the fences of their national languages and collaborate. Scholars of the classical languages with a long and interrelated history (Greek, Latin) have set an example and could play a leading role in this.

- **Extent to which morphological and syntactic information needs to be built into lexical representations:** when tagging texts, a first source of information about the parts of speech of tokens are lexica. A very obvious pitfall here concerns the distinction of the lexical *Part of Speech* (PoS) of a wordform – normally stored in the lexicon – and the syntactic PoS of a token of that form in a sentence. Obviously, these two assignments need to be distinguished. On the one hand, the PoS of a wordform stored in the lexicon can be used as a reference when tagging sentences in order to reduce the number of unknown tokens (obviously, this reduction supports any statistical tagging). On the other hand, the tagger may need to overwrite the lexicon information.

Take the example of past participles, which in the lexicon are normally subsumed under a corresponding verb lemma: in German, for instance, participles can be used to derive adjectives (like in ‘Der zerbrochene Krug’/‘The Broken Jug’) which have to be tagged appropriately. Thus, one has to balance the information taken from the lexicon against what has to be overwritten by the tagger. A way out of this problem is to tag both kinds of PoS: the lexical and the syntactic one. In any event, derivational knowledge (e. g. about the derivation of adjectives from participles) has to be included in the lexicon so that the search space of the tagger can be reduced. Given, for example, a verb like *lesen* (‘to read’) in German, a large set of nouns can be derived from it: *der Leser, der Lesende, das Gelesene, die Lesbarkeit, die Leserei, die Lesung* etc. Thus, one should not underestimate the additional amount of information to be stored in the lexicon if one has to consider, for example, a set of 20,000 verbs of a language. Derivational knowledge is morphological knowledge that is included here in the lexicon in order to guide the tagging of syntactic information in sentences.

Note that the range of ‘syntactically motivated’ PoS can be much larger than what is distinguished lexically in the lexicon. Take the example of conjunctions where one can distinguish between subordinating conjunctions (subjunctions) and coordinating ones. Obviously, dependency parsing can be boosted by making this distinction during PoS-tagging. Thus, morphological or lexical ontologies of parts of speech can depart from

their syntactically motivated ones. Relying on some ‘universal’ PoS tagsets (Petrov et al 2012) does not solve this task. Once more, the reason is simply that if we want to reduce error rates of text parsing we need to include more and more information in the lexicon, that is, we need to make more distinctions, distinctions that are abstracted by universal tagsets or universal rule sets Marneffe et al (2014). In other words: while universal tag- or rule sets aim at the interoperability or comparability of methods, the humanities need in many cases rather fine-grained models that map the specifics of a given language or corpus and, thus, contrast with interoperability.

Another obvious example in favor of including syntactic information in the lexicon relates to the valency of verbs. Knowledge about this valency can guide dependency parsing and corresponding disambiguation processes (when distinguishing, for example, between complements of verbs and nouns). Once more, the amount of information to be considered here is enormous – it is even higher if we consider the requirement to account for variation of this information over time (see below). However, in order to meet the very low error rates acceptable to humanities scholars, there seems to be no alternative to more ambitious projects of building lexica.

To be more precise: any decision about what to include in the lexicon hinges upon the need to reduce error rates of tagging (historical) texts for humanities scholars. In this line of thinking, we always get a reason to extend the lexicon as much as possible: given the plethora of annotation desired by scholars (and not just from the point of view of NLP), most of the relevant information units still cannot be tagged automatically. Thus, it is desirable to put as much information as possible into the lexicon in order to make tagging less error prone. From this point of view, present-day full-form lexica (although usually including information about PoS and inflectional paradigms) are insufficient.

This approach may further interdisciplinary collaboration between computer scientists and scholars in the development of lexica and taggers based thereon. An example of such an interdisciplinary project is reported in (Mehler et al 2015) where historians work together with computational linguists in the exploration of Latin texts. The project deals with the genre- and register-related classification of medieval Latin texts based on their pre-processing in terms of lemmatization and PoS-tagging. To this end, the authors developed a large-scale Latin lexicon (primarily based on inflection patterns that produce around 11 million wordforms out of ca. 250,000 lemmata). The lexicon is used as a reference for PoS-tagging whenever it lists a single PoS for a form. Beyond that, tagging is done by means of a CRF that is superimposed by a set of short-scale ‘syntactic’ rules. In this sense, a hybrid approach is followed where lexical information is combined with a syntactic knowledge base and a statistical tagger. One should not underestimate the amount of work entailed by an approach in which the syntactic rules are handcrafted as are many of the patterns and even entries of the underlying lexicon. As it stands, such a labor-intensive approach (somehow reminiscent of human computation) is indispensable for text processing and, thus, for the generation of classification results acceptable to historians.

Several tasks undertaken by the Herodotos Project for Ethnohistory (Ohio State University/Ghent University) illustrate the necessary interplay of human correction with machine generation of data. These include: the determination of error rate and causes of error in the application of the Stanford Classifier to the identification of group names in the English texts of the Perseus corpus of ancient authors; the refinement of the classifier to deal with authors of different genres and periods; the development of Latin and Greek language classifiers suitable for identifying group names; automated XML markup of

the texts for which we have complete lists of (edited/corrected) group names, using the *TTLab Latin Tagger*⁶ (TLT) for Latin and XML-/TEI-based tagging (Mehler et al 2015), and marking up Perseus code for group names by an automated process.

- **Relationship between dictionaries and chronology:** A key role for finding information about the history of a word and its usage is played by the dating of its sources in historical dictionaries. Changes of spelling and morphology of a lexical item, the first record of its use and meaning etc. are usually documented in the great national dictionaries (OED, DWB², TLF etc.). The definitions of a lemma connected to a certain language stage could also be looked up in a period specific lexicon. Of special interest in searching historical texts for definitions are borrowings, especially from Latin into German, English and other European vernaculars. The Latin borrowings e. g. of German from its first recordings in bilingual word lists in the 8th century through all the following periods up to the 19th century are immense. In religious texts of the Middle Ages as well as in scientific writings of today there is hardly any sentence without Latin traces. Latin loans in German books printed from about 1500 were marked for a long time by a change of fonts: Fractura had been used for German, antiqua for Latin.

In lemmatizing historical texts the change of fonts is essential in order to filter out loans and find the appropriate time related definitions in the *Deutsches Fremdwörterbuch* (Schulz et al. 1995), the sister of the DWB which from its inception was not meant to contain loanwords (*Fremdwörter*). The borrowings from Latin consist not only of loanwords, which are usually taken over together with a specific meaning (see the definitions in the national dictionaries to Medieval Latin), but also of loan translations (e. g. Latin *re-sur-rect-io* and its morpheme based rendering in German *Auf-er-steh-ung* which goes back to MHG *ûf-er-stand-unge*). Translating key Christian terms in the Middle Ages led to a variety of synonyms of which in the course of time often only one has survived (of six synonyms for Latin *gratia* with its specific Christian meaning in OHG only one made it into MHG *genâde*, NHG Gnade). For determining which concept is represented by which lemma and definition we need a semantic index to the historical dictionaries. This is a real challenge for digital humanists trying to explore the lexical history of an expression and its definitions through time and place. An inspiring example is the *Historical Thesaurus of the OED* by Christian Kay which has been integrated into the *OED online*.

It is a commonplace that the meaning of a lexeme changes over time. However, it does not do so according to a single timescale. Thus, by analogy to Domingos (2008), we may speak of modeling the variation of lexical items in terms of *structured time* in order to account, for instance, for different processes of temporal variation (e. g., function words change according to a longer timescale than content words). This variation can be conditioned by the dynamics of genres and registers in which the lexical items are preferably used (Halliday 1977, Halliday 1991). In such cases, models of genres and registers are additionally required. Thus, beyond morphological and syntactical knowledge we may also include pragmatic knowledge in the lexicon. Time is just a gateway for this kind of knowledge.

Thus, a central challenge of automatic, lexicon-based text analysis of historical texts concerns the requirement to cover time as a constitutive parameter of lexicon formation. That is, the variation of the morpho-syntactic realizations of lexical items over time have to be considered as an integral part of the lexeme/syntactic word/wordform relation. So far, little is done in this respect: either the lexica do not contain information about lexical

⁶ See <http://prepro.hucompute.org> and <http://collex.hucompute.org>.

variation (applying, for example, lexica of classical Latin to medieval Latin texts) or the taggers do not operate in a time-sensitive manner. In order to understand possible pitfalls of the latter case consider the task of tagging multilingual texts: taggers are typically language-specific; if an input text of language *A* contains text spans (e.g. citations) of another language, the tagger tries to tag these spans as instances of language *A* – obviously, this is an erroneous procedure. Rather, what should happen here is that the tagger starts with language detection for any text span in order to select the corresponding language-specific tagger for it. The same should happen along the time axis where time period-sensitive taggers are selected to tag corpora of historical texts that instantiate several stages in the development of one or more languages. As it stands, current taggers are not powerful enough to account for such requirements of stratified tagging – stratified with respect to time, language, register, genre etc.

- **Linking lexica via *hyperlemmata*:** above, we argued that rather than abstract tagsets, fine-grained lexicon models are needed to meet the requirements of, say, philologists, who look for the specifics of certain texts rather than for a generalized model, say, of the PoS realized by them. Such an approach runs the risk of adapting its lexicon model to the specifics of the underlying corpus in such a way that interoperability of methods and comparability of findings is negatively affected. In order to provide a way out of this fallacy, we may think of using *hyperlemmata* to establish links between the lemmata of different lexica. This model is in line with approaches like Petrov et al (2012) and Marneffe et al (2014), but with a focus on lexemes instead of PoS or dependency rules. Given a unified lexicon model based on hyperlemmata one can envision ‘translations’ between different lemmatizations of the same text. Alternatively, one can envision abstract search queries based on hyperlemmata that are automatically mapped onto the specifics of the underlying lexica. Such an additional layer of modeling lexica entails a further level of labor-intensive research. However there seems to be no alternative to such an approach if our goal is to switch between different lexical ontologies.
- **Compiling lexica automatically (definition generation):** since processing historical languages is reminiscent of processing low-resourced languages in that it faces related challenges, it is necessary to think of standardized procedures for the rapid, less error prone compilation of lexica even out of (small) corpora of historical texts. Here, we envision a combination of methods of (i) computational linguistics for learning, for example, inflection patterns, valency patterns or word-order patterns, (ii) text-technological methods of building and maintaining lexical databases and (iii) methods of human computation for the fine-grained adaptation and extension of the resulting lexica. On the basis of such a procedure, one can envision an application that allows for estimating the complexity of building a lexicon for a given historical language starting from a given corpus of a certain size. Such an application could help interdisciplinary projects distribute the various tasks of compiling the lexicon among project members.

4.4.3 Computational Analysis of Literary Texts

In addition to the structured information in human- and machine-readable lexica, computational linguists and digital humanists work with increasingly large bodies of unstructured text. To speak very broadly, this text varies greatly in the specificity of its metadata, the consistency of its editing, and the standards and accuracy of its transcription. On the one hand, creators of lexica and other linguistic resources have always used corpora to investigate and illustrate linguistic facts, and textual critics have always been concerned with the basis of our knowledge of texts. On the other hand, the wide availability of electronic texts and

means for their automatic analysis encourage us to think more systematically about the interplay of lexicon and corpus.

We believe, therefore, that important research questions will continue to center around our ability to augment structured resources such as lexica with inferences from unstructured text and how to exploit lexica to improve automatic processing. Among the specific problems we discussed were:

- practical problems in compiling corpora to work with, in particular for long-term diachronic analysis including multiple language stages, typefaces (e. g., Fractura), and genres;
- constructing corpus-specific lexica and refining existing lexica with corpus data;
- adapting standard NLP tools to domains (e. g., literary texts) that may be divergent from the newspaper texts on which they were trained;
- interpreting automatic clustering methods such as topic modeling extraction from texts: the intersection of computational analysis of text and the lexicon, since here word-meanings make a difference;
- automated thematic analysis;
- automated plot summaries; and
- computer-aided stylistic analysis.

For example, one problem in applying topic models and related approaches to historical texts is that any semantic analysis should not only consider wordforms, but rather lexemes or – better – lexeme groups (*Lexemverbände* in German) which subsume lexemes based on the same stem even if they belong to different part-of-speech classes (an example is *fliegen*, *Flug*, *Flieger* etc.). In order to do this, a very good lemmatization is needed. As discussed above, this is a task that is not completely solved in the case of historical texts. Here, we still need to do a lot even in terms of lexicon building. However, presentations of clouds of wordforms subsumed under the same topics will hardly convince philologists or historians who – as outlined above – expect very low error rates. A wordform is a formal unit and not a semantic unit. Lexemes in the lexicon are dually articulated in the sense of de Saussure: formally by the wordforms by which they are realized and semantically by the meaning that they carry. If topic modeling aims at drawing level with this view, it should be combined with a very thorough pre-processing of historical texts – beyond what is currently done in many approaches to topic modeling. Here, historians, philologists and computational linguists should go hand in hand in order to further develop their methods – possibly by example of topic models that are well established on the ground of taggers as described in section 4.4.2. This can be a way out of detecting, for example, function words as part of word clouds attributed to a certain topic, where these function words do not occur in the cloud because of carrying a certain meaning, but just due to the statistics of the given text.

What kind of structure of the lexicon would enable a better analysis of literary texts? Strategies to improve text analysis, which informations of a digital lexicon can be employed (for example hypernymy/hyponymy)?

In order to better meet the requirements of text analysis with the help of lexica, the following information objects should be included into lexicon formation: the lexicon should cover derivation relations in order to allow for modeling lexeme groups (see above). It must consider time as an attribute of any relation and any attribution in the lexicon (e. g., Which lexeme is realized during which period by which wordform carrying which grammatical information? etc.). Beyond time, each lexicon entry should be equipped with an expressive attribute model that allows for mapping various syntactic, semantic, pragmatic, genre- or register-specific information units (e. g., sentiment/polarity, connotations, semantic classes (e. g., anthroponyms, oikonyms, chrononyms etc.)).

4.4.4 Error Detection and Correction

An important issue with any application of computational methods to text is the degree to which errors occur in the automatic processing. While scholars in some areas of computer applications may be satisfied with a small error rate (say 1%, for optical character recognition of documents printed within the last hundred years or so), humanists tend to be very concerned about the integrity of the text that they are working with, and tend to express great dissatisfaction with even tiny error rates smaller than 1%. Thus, it is a concern to be able to detect errors, to predict the rate at which they are likely to occur, to characterize their effects on subsequent processing, and to be able to do something about the errors if possible. Among the applications we considered that could generate errors (while at the same time, of course, generating electronic output that is very useful and usable) was optical character recognition to create electronically manipulable texts.

- classification of errors (OCR errors, lemmatization errors, POS-tagging errors, parsing errors etc.);
- consequences in statistical analysis;
- how error rates affect results;
- the extent to which errors are random or patterned; and
- understanding the impact of errors in functions and propagation of errors relationship of dictionaries, functions and errors.

A central challenge of any error analysis concerns the availability of online tools for comparative error analyses by which errors can be classified and displayed in terms of summaries (e. g., by decreasing frequency). This is needed to allow for a more rapid and comprehensive detection and elimination of errors. Current systems either only provide summary data (in the form of F-measure statistics) or only selected error analyses by discussing some use cases. However, what is needed is a systematic overview of the whole range of errors being made by automatic text analysis, an overview that human users can use to guide future processes of text analysis in order to guarantee lower error rates. To this end, computational linguists building annotation tools, digital humanists (providing web-based interfaces for the usage of these tools), and humanities scholars should cooperate much closer in order to meet the low-error-rate requirement of the humanities.

4.4.5 Conclusions and Future Research

Digital linguistic resources such as lexica are necessary for making progress in many areas of natural language processing; moreover, the availability of digitized corpora and automatic annotation methods can make creating these resources a collaborative effort between linguists, philologists, and computer scientists. We see several opportunities for strengthening these collaborations, for creating new linguistics resources, and for analyzing and mitigating the errors in human and computational annotation processes.

Acknowledgement. The group thanks Bettina Berendt for her fruitful hints, comments, and discussion of the topics discussed by this working group.

References

- 1 Gregory Crane. Building a digital library: the perseus project as a case study in the humanities. In *Proceedings of the first ACM International Conference on Digital libraries*, DL'96, pages 3–10, New York, NY, USA, 1996. ACM.
- 2 Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal stanford dependencies: A

- cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 4585–4592, 2014.
- 3 Pedro Domingos. Structured machine learning: Ten problems for the next ten years. *Machine Learning*, 73:3–23, 2008.
 - 4 Johannes Fournier. New directions in middle high german lexicography: Dictionaries inter-linked electronically. *Literary and Linguistic Computing*, 16:99–111, 2001.
 - 5 Kurt Gärtner. The new middle high german dictionary and its predecessors as an interlinked compound of lexicographical resources. In *Digital Humanities 2008, Oulu, Finland, Book of Abstracts*, pages 122–124, 2008.
 - 6 Michael A. K. Halliday. Text as semiotic choice in social context. In Teun A. van Dijk and J. S. Petöfi, editors, *Grammars and Descriptions*, pages 176–225. De Gruyter, Berlin/New York, 1977.
 - 7 Michael A. K. Halliday. Towards probabilistic interpretations. In Eija Ventola, editor, *Functional and Systemic Linguistics*, pages 39–61. De Gruyter, Berlin/New York, 1991.
 - 8 Hans Schulz et al. *Deutsches Fremdwörterbuch, begonnen v. Hans Schulz, fortgeführt v. Otto Basler, weitergeführt im Institut für deutsche Sprache, Bd. 1–2 [A–Pyramide], Straßburg 1913 und 1942, Bd. 3–7 [Q bearb. v. Otto Basler, P–T bearb. v. Alan Kirkness, U–Z bearb. v. Gabriele Hoppe; Bd. 7 Quellenverzeichnis, Wortregister, Nachwort hg. v. Alan Kirkness], Berlin 1977–1988. – Deutsches Fremdwörterbuch, 2. Auflage, völlig neunearb. im Institut für deutsche Sprache, Bd. 1ff. (bisher Bd. 1–7 [A–Präfix – hysterisch] bearb. v. Gerhard Strauß u.a., volume 1. de Gruyter, Berlin/New York, 1995ff.*
 - 9 Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, and Ladislav Zgusta, editors. *Wörterbücher. Ein Internationales Handbuch zur Lexikographie. 3 Teilbände. Handbücher zur Sprach- und Kommunikationswissenschaft 5, 1; 5, 2; 5, 3.* de Gruyter, Berlin / New York, 1989; 1990; 1991.
 - 10 Henry George Liddell and Robert Scott. *A Greek-English lexicon: With a revised supplement.* Clarendon Press, Oxford, 1996.
 - 11 Alexander Mehler, Tim von der Brück, Rüdiger Gleim, and Tim Geelhaar. Towards a network model of the coreness of texts: An experiment in classifying Latin texts using the tlab latin tagger. In Chris Biemann and Alexander Mehler, editors, *Text Mining: From Ontology Learning to Automated text Processing Applications*, Theory and Applications of Natural Language Processing. Springer, Berlin/New York, 2015. To appear.
 - 12 Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC 2012*, pages 2089–2096, Istanbul, Turkey, 2012.

5 Panel Discussions

5.1 Evolving Computation, New Research Directions and Citizen Science for Ancient Greek and the Humanities

Gregory R. Crane (Tufts University, US)

License  Creative Commons BY 3.0 Unported license
© Gregory R. Crane

URL <http://sites.tufts.edu/perseusupdates/2014/09/29/opening-up-classics-and-the-humanities-computation-the-homer-multitext-project-and-citizen-science/>

Increasingly powerful computational methods are important for humanists not simply because they make it possible to ask new research questions but especially because computation

makes it both possible – and arguably essential – to transform the relationship between humanities research and society, opening up a range of possibilities for student contributions and citizen science: <http://homermultitext.blogspot.de/>, <http://www.homermultitext.org/>.

My departure point for this paper were questions during a workshop on Computational Humanities, at Schloss Dagstuhl in July 2014 that Manfred Thaller posed to me in response after a talk I delivered, and then as part of a podium discussion. I had argued that we needed to advance research projects that provided our students with an opportunity to make substantive contributions to research as a central part of their education. In so doing, I echoed Wilhelm von Humboldt, who argued that students in a university should always be engaged in advancing human understanding – mastering set curriculum was for primary and secondary school. For Humboldt, the challenge of new questions and interests from students was one of the great intellectual advantages of working in a university rather than in a research institute. Manfred reminded me that if we focused too much upon serving the students, then we would do a disservice to research.

The challenge is to establish a productive tension between the established interests of the faculty and the fresh questions of the students. I conclude this paper by describing the Homer Multitext Project and how, in opening up new opportunities for student contributions and research, this project shifted my idea of how the field should move forward. Research should not solely serve students (then it runs the risk of being dumbed down) but it should not only serve specialist researchers (then it runs the risk of becoming detached scholasticism). In the ideal case, we identify research that challenges (and captivates) the most advanced researchers but that also engages a broader audience and provides multiple opportunities for contribution. In the study of Greek and Latin, I see this productive tension leading to re-organization (and, in the US at least, a revival) of very traditional philological questions with very new methods and in a much more decentralized and collaborative culture.

This paper thus argues that the most important consequence of computation for the Humanities lies not in the new research questions that are now appearing but in the fact that research with increasingly large bodies of digitized source materials opens up opportunities, indeed the necessity, for a new, more open culture of intellectual production, one that is less hierarchical, more focused on collaborative inquiry, more dynamic, and, in my view, more effective as an environment for broad and deep intellectual development. Computation allows – perhaps more accurately, challenges – humanists to redefine their relationship to their students and to society as a whole.

A link to the full document can be found at <http://tinyurl.com/kner5dk>.

5.2 The Humanities are about research, first and foremost; their interaction with Computer Science should be too.

Manfred Thaller (University of Cologne, Germany)

License  Creative Commons BY 3.0 Unported license
© Manfred Thaller

This statement describes with the privilege of hindsight a view on a controversy about the relative weight to be assigned to various priorities which the application of computer science to Humanities research questions should support. The controversy was started by my response to a statement of Greg Crane about the importance of pursuing educational goals by computer supported systems in the Humanities. I argued, that the application of

Computer Science to the Humanities at the university level must gain its intellectual merit primarily from the contribution it makes to the research agenda of the Humanities. The controversy about the relative importance of these two goals can be quite sharp, depending on the implications one of these two priorities has within a political argument for the way in which Computer Science and the Humanities are to cooperate in a given university context. Obviously, when such a political context does not exist, the two goals are not contradictory as such. Nevertheless, the following theses shall summarize and clarify my appeal for a strong focus on the research agenda in the Humanities to be supported by Computer Science.

Assumption 1.1: The interest of the Humanities in Computer Science has, from the very beginning, been directed by two goals, which can easily be mixed up with confusing results. Humanities research requires many routine tasks, which are plain drudgery: Busa's dream of never have to go through Aquinas line by line any more to look for the forms of the root of a specific word, is the obvious example. On the other hand, there has always been the promise that some methods supported by Computer Science might open up the way to explanations for Humanities phenomena, which open up new epistemic vistas for the disciplines producing these explanations.

Assumption 1.2: While the interest in both aspects of this interdisciplinary field is perennial, the relative emphasis assigned to them depends intellectually on the state of the epistemological discussion in the Humanities at large or their individual disciplines. Humanists who are impressed by C. P. Snow or K. R. Popper assign different priorities to interdisciplinarity, than those following P. K. Feyerabend.

Assumption 2.1: The Humanities consider themselves currently in a crisis, or rather, primarily the Anglo-American branch of the Humanities does. In those countries they react by an intensive discussion on how low-level applications of technology can make the disciplines look more modern and relevant. As this is a defensive movement, it frequently reacts frightened against applications which require a more thorough understanding of Computer Science, which is seen as additional competition, which is more likely to sharpen the crisis of the Humanities, rather than alleviate them.

Assumption 2.2: This I consider a dangerous fallacy. The Humanities do not become more respectable by showing that they employ the same simple tools which everybody else uses in the year 2014. Nor does the ability to teach critical thinking at the level of the gymnasium justify a research agenda.

Assumption 2.3: The Humanities need a broad vision, why they are so fascinating, that society at large should support them. This is not done by inventing short term economic benefits, but by presenting goals which have such a wide appeal, that society is willing to support them, even if no short term benefits are generated. Hubble does not lower unemployment rates; it promises to unveil fascinating secrets.

Assumption 3.1: Computer Science for a long time has been a discipline which emphasized the necessity of data being highly structured and free of contradictions as a precondition for their processing, even if in some rather exotic branches a theoretical interest in applications vulnerating these preconditions has always been existing. As it has changed from a discipline supporting individual solutions for specific problems into the conceptual and theoretical backbone of an integrated infosphere, Computer Science can less and less define what properties the data should have, it intends to process and has therefore to handle whatsoever comes along.

Assumption 3.2: The Humanities have since their earliest inception always been focusing on the ability to draw a maximum of conclusions from a rather limited amount of information, they could access physically. The only start to notice, that this barrier has broken down. The primary qualification of a Humanities' researcher of the year 2050 will not be, how to lovingly extract insights from a few isolated bits of information, but how to meaningfully integrate the information contained in the largest possible set of data.

Assumption 3.3: There is a convergence, therefore, between the approach towards data to be supported by Computer Science and the Humanities.

Thesis 1: To reach the vision postulated as necessary by assumption 2.3, the Humanities have to focus more strongly again at the epistemic implications of methods which can be supported algorithmically (cf. assumption 1.1 / 1.2), as only so the challenge posed by assumption 3.2 can be answered successfully.

Thesis 2: Care has to be taken, that the high visibility of low level approaches to “Digital Humanities” described in assumptions 2.1 and 2.2 do not obscure the developments needed for the support of thesis 1.

Thesis 3: To support thesis 1, a joined research agenda between the Humanities and Computer Science is necessary, which does not restrict itself to the application of known algorithmic solutions on the knowledge domain of the Humanities, but uses the challenges described in assumption 3.2 to help solving the challenges described in assumption 3.1. Realizing, that is, what has been postulated by assumption 3.3.

Participants

- Loretta Auvil
University of Illinois at Urbana
Champaign, US
- David Bamman
Carnegie Mellon University, US
- Sven Banisch
Universität Bielefeld, DE
- Chris Biemann
TU Darmstadt, DE
- Christopher Brown
Ohio State University, US
- Marco Büchler
Georg-August-Universität
Göttingen, DE
- Gregory R. Crane
Tufts University, US
- Jana Diesner
University of Illinois at Urbana
Champaign, US
- Christiane D. Fellbaum
Princeton University, US
- Anette Frank
Universität Heidelberg, DE
- Kurt Gärtner
Universität Trier, DE
- Siegfried Handschuh
Universität Passau, DE
- Gerhard Heyer
Universität Leipzig, DE
- Fotis Jannidis
Universität Würzburg, DE
- Brian Joseph
Ohio State University, US
- Cathleen Kantner
Universität Stuttgart, DE
- Kai-Uwe Kühnberger
Universität Osnabrück, DE
- Jonas Kuhn
Universität Stuttgart, DE
- Andy Lücking
Universität Frankfurt, DE
- Alexander Mehler
Goethe-Universität Frankfurt am
Main, DE
- David Mimno
Cornell University, US
- Meinard Müller
Univ. Erlangen-Nürnberg, DE
- Andrea Rapp
TU Darmstadt, DE
- Geoffry Rockwell
University of Alberta, CA
- Szymon Rusinkiewicz
Princeton University, US
- Joachim Scharloth
TU Dresden, DE
- Maximilian Schich
The Univ. of Texas – Dallas, US
- Ute Schmid
Universität Bamberg, DE
- Susan Schreibman
NUI Maynooth, IE
- David A. Smith
Northeastern University –
Boston, US
- Caroline Sporleder
Universität Trier, DE
- Wolfgang Stille
Universitäts- und
Landesbibliothek Darmstadt, DE
- Manfred Thaller
Universität Köln, DE
- Claire Warwick
University College London, GB
- Katharina A. Zweig
TU Kaiserslautern, DE



Digital Palaeography: New Machines and Old Texts

Edited by

Tal Hassner¹, Robert Sablatnig², Dominique Stutzmann³, and
Ségolène Tarte⁴

1 Open University of Israel – Raanana, IL, hassner@openu.ac.il

2 TU Wien, AT, sab@caa.tuwien.ac.at

3 Institut de Recherche et d'Histoire des Textes (CNRS) – Paris, FR,
dominique.stutzmann@irht.cnrs.fr

4 University of Oxford, GB, segolene.tarte@classics.ox.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 14302 “Digital Palaeography: New Machines and Old Texts”, which focused on the interaction of Palaeography and computerized tools developed in Computer Vision for the analysis of digital images. This seminar intertwined research reports from the most advanced teams in the field and interdisciplinary discussions on the potentials and limitations of future research and the establishment of a community of practice in Digital Palaeography. It resulted in new research directions in the Computer Sciences and new research strategies in Palaeography and in a better understanding of how to conduct interdisciplinary research across all the fields of expertise involved in Digital Palaeography.

Seminar July 20–24, 2014 – <http://www.dagstuhl.de/14302>

1998 ACM Subject Classification I.7 Document and Text Processing, H.1.2 User/Machine Systems, D.2.1 Requirements/Specifications, H.3.3 Information Search and Retrieval, H.5.2 User Interfaces, I.4 Image Processing and Computer Vision, I.5 Pattern Recognition, J.5 Arts and Humanities

Keywords and phrases Handwriting Recognition, Interdisciplinarity, Epistemology, Middle Ages, Manuscript studies, Expertise, Knowledge exchange

Digital Object Identifier 10.4230/DagRep.4.7.112

1 Executive Summary

Dominique Stutzmann

Ségolène Tarte

License © Creative Commons BY 3.0 Unported license
© Dominique Stutzmann and Ségolène Tarte

Digital Palaeography emerged as a research community in the late 2000s. Following a successful Dagstuhl Perspectives Workshop on Computation and Palaeography (12382)¹, this seminar focused on the interaction of Palaeography and computerized tools developed in Computer Vision for the analysis of digital images. Given the present techniques developed to enhance damaged documents, optical text recognition or computer-assisted transcription, identification and categorisation of scripts and scribes, the current technical challenge is

¹ <http://dx.doi.org/10.4230/DagMan.2.1.14>



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

Digital Palaeography: New Machines and Old Texts, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 112–134

Editors: Tal Hassner, Robert Sablatnig, Dominique Stutzmann, and Ségolène Tarte



Dagstuhl Reports

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

to develop “new machines”, i. e. efficient solutions for palaeographic tasks, and to provide scholars with quantitative evidence towards palaeographical arguments, even beyond the reading of “old texts” (ancient, medieval and early modern documents), which is of interest to the industry, to the wider public, and to the broad community of genealogists.

The identified core issue was to create the conditions of a fluid and seamless communication between Humanities and Computer Sciences scholars in order to advance research in Palaeography, Manuscript Studies and History, on the one hand, and in Computer Vision, Semantic Technologies, Image Processing, and Human Computer Interaction (HCI) systems on the other hand. Indeed, researchers must articulate their respective systems of proof, in order to produce efficient systems that present palaeographical data quickly and easily, and in a way that scholars can understand, evaluate, and trust. To establish fruitful collaborations, it is thus essential to address the “black box” issue, to make a better use of the outreach potential offered by computerized technologies to enrich palaeographical knowledge, and to facilitate the sharing of both the CS and palaeographical methodologies.

This seminar was able to shed light onto two major evolutions between 2012 and 2014; these notable shifts are to do with interdisciplinary communication and with access to “black box” expertise. On the one hand, the notion of “communication” or “bridging the gap” (as expressed by seminar 14301, which took place in conjunction with our own seminar) has become more specific in that issues and problems are now better identified, understood, and expressed. While the two-fold expression “digital palaeography” might lead one to believe that the communication involves only two sorts of actors, it has been expressed in ways clearer than ever that Digital Palaeography as a field is much more complex than a simplistic adjunction of Computer Sciences and Palaeography; indeed CS research, engineering and software development, support and service, linguistics, palaeography, art history, and cultural heritage institutions (Galleries, Libraries, Archives, and Museums – GLAM) all form part of the Digital Palaeography research arena. Good communication requires correct identification of the roles and competence of each actor, and a well-balanced project has to associate/include/foresee the participation of the other actors. It is for example important to clarify that palaeographers are not responsible for copyright or image quality provided by GLAM institutions, in the same way as CS researcher are not responsible for designing interfaces. Within each community, a better understanding of methods and interests of the actors of the other communities is needed to find the right partners (e. g.: keyword spotting is not alignment; writer identification is not script classification). On the other hand, the “black box” issue seems to have been addressed by most teams through the introduction or increase of interactivity of the software tools they presented; interactivity was used not only as a means to produce clear and convincing results, but also to overcome the shortcomings of strictly automatic approaches. In this sense, the reintroduction of “the human into the loop” (or “the use of the users”) is part of a process allowing a better understanding on both sides. The “human in the loop” can and should be integrated at all stages, and, even if this need is not always perceived, it is crucial that substantial efforts be dedicated to making implicit assumptions or knowledge explicit. Special attention should be given to avoid the development of tools relying on tautological approaches where tools or datasets incorporate expectations as an underlying (and often implicit) model. In this regard, one cannot overestimate that an unclear result is as important for historians as a clear-cut clustering. In the middle, the “human” gives feedback on preliminary results, enables the enhancement and improvement of the model, as well as creates ground-truth. The display of intermediary results and the integration of user feedback within the process are a welcome solution offered by the latest developments. Likewise, palaeographers have developed new

strategies, in their ways of formulating tool requirements or expressing requirements for which they can evaluate the results themselves, regardless of the software being an opaque black-box (P. Stokes, D. Stutzmann, M. Lawo with B. Gottfried).

Overall, this seminar seems to have operated a paradigm shift from black-box issues to trust issues, in the sense that when we first identified black-box issues, we focussed on “computational black boxes”, when “human black boxes” are in fact just as problematic. Instead of focussing on computational black-boxes as an issue, we were able to formulate that the important endeavour is that of establishing trust in the respective methodological approaches to the research questions of the research domains. This trust in methodologies is usually mediated by human interactions (“humans in the loop” again!), and the ways in which scholars are able to share an intuitive understanding of their respective expertises with non-experts.

It hence follows that a new (technical) challenge arises, consisting in the creation and implementation of an integrated software tool, web service suite, or environment that would allow users to access and work with extant datasets and tools. The impetus to take up this challenge resides as much in the Humanities as it does in the Computer Sciences. By aggregating the multiple, isolated, specific tools developed by CS researchers through a common access point, digital humanists would support the development of better evaluation metrics and promote a wider use of CS technologies among more traditional Humanities scholars, who could thus become more aware of the existing tools, more autonomous (i. e. less dependant on CS researchers) and thereby empowered. As a reciprocal positive effect, CS researchers could more easily validate their results and gain access to a wider range of annotated datasets. This challenge is also naturally related to trending key concepts such as “interoperability” and “open access”. It furthermore engages with the question of the nature of success metrics in the Humanities, where a successful tool is not only the one giving the best results, it is also one enjoying wide acceptance and a large number of users. Improving ergonomics is mandatory, to put the user in the middle and to accumulate a consistent critical mass of annotations (both as feedback and ground-truth).

2 Table of Contents

Executive Summary

<i>Dominique Stutzmann and Ségolène Tarte</i>	112
---	-----

Overview of Talks

Interdisciplinary Approach to the Study of Tibetan Manuscripts and Xylographs: The State of the Art and Future Prospects <i>Orna Almogí</i>	117
Encoding Scribe Variability <i>Vincent Christlein</i>	117
Algorithmic Paleography <i>Nachum Dershowitz</i>	118
Appearance Modeling for Handwriting Recognition <i>Gernot Fink</i>	118
Separating glyphs of handwritings with Diptychon <i>Björn Gottfried</i>	120
Deciphering and Mapping the Socio-Cultural Landscape of 12th Century Jerusalem: Texts, Artifacts and Digital Tools <i>Anna Gutgarts-Weinberger and Iris Shagrir</i>	120
Positioning computational tools <i>Tal Hassner</i>	122
DIVADIA & HisDoc 2.0 Approaches at the University of Fribourg to Digital Paleography <i>Marcus Liwicki</i>	123
Word spotting in historical manuscripts. The “Five Centuries of Marriages” project <i>Josep Lladós</i>	124
Modern Technologies for Manuscript Research <i>Robert Sablatnig</i>	124
tranScriptorium <i>Joan Andreu Sanchez Peiro</i>	126
Text Classification and Medieval Literary Genres <i>Wendy Scase</i>	126
Describing Handwriting – Again <i>Peter A. Stokes</i>	127
Bridging the gap between Digital Palaeography and Computational Humanities <i>Dominique Stutzmann</i>	128
Digital Palaeography. Text-Image Alignment and Script/Scribal Variability (ANR ORIFLAMMS / Cap Digital) <i>Dominique Stutzmann</i>	129
Digital Images of Ancient Textual Artefacts: Connecting Computational Processing and Cognitive Processes <i>Ségolène Tarte</i>	130

116 14302 – Digital Palaeography: New Machines and Old Texts

Text classification	
<i>Nicole Vincent</i>	131
Diplomatics and Digital Palaeography	
<i>Georg Vogeler</i>	132
A Graphical Representation of the Discussed Subjects	132
Participants	134

3 Overview of Talks

3.1 Interdisciplinary Approach to the Study of Tibetan Manuscripts and Xylographs: The State of the Art and Future Prospects

Orna Almogi (Universität Hamburg, DE)

License © Creative Commons BY 3.0 Unported license
© Orna Almogi

From the point of view of a student of the history of ideas who is primarily interested in the intellectual culture, intellectual history, philosophy, and religion of any given civilization, past and present, it is assumed that there is no effective way to gain a nuanced and well-founded knowledge of them without a profound knowledge of the pertinent languages and without extensively exploring the diverse indigenous textual sources. Despite significant progress that has been made during the past decades in this regard in the field of Classical Tibetan Studies, a relatively new discipline, scholars have barely managed to scratch the surface of the enormously vast, diverse, and rich textual material that has come down to us in the form of manuscripts and xylographs produced by the Tibetan civilization over the centuries. Recent decades have witnessed a significant increase in the accessibility of old Tibetan (mainly Buddhist) texts produced and transmitted from the seventh century until the present. These new discoveries of old primary textual material have no doubt significant implications in the field, posing new challenges and at the same time offering fascinating new opportunities for Tibetologists. However, this tremendous increase in the accessibility of hitherto inaccessible and unexplored textual material some of it fragmentary and often no longer in its original place of deposit but scattered over various libraries around the world heighten the desire to refine existing research tools and seek new ones that are more efficient and more powerful for investigating this material and the ideas transmitted therein. In my presentation I presented the state of affairs in the field of Tibetan textual studies, briefly discussing the major difficulties Tibetologists face in dealing with the large and diverse textual material, and finally described three computerized tools aiming at facilitating Tibetan textual studies that are currently in development.

3.2 Encoding Scribe Variability

Vincent Christlein (Universität Erlangen-Nürnberg, DE)

License © Creative Commons BY 3.0 Unported license
© Vincent Christlein

Joint work of Christlein, Vincent; Bernecker, David; Hönig, Florian; Angelopoulou, Elli
Main reference V. Christlein, D. Bernecker, F. Hönig, E. Angelopoulou, “Writer identification and verification using GMM supervectors,” in Proc. of the 2014 IEEE Winter Conf. on Applications of Computer Vision (WACV’14), pp. 998–1005, IEEE, 2014.
URL <http://dx.doi.org/10.1109/WACV.2014.6835995>

Like faces or speech, handwritten text can serve as a biometric identifier. This talk gives an overview of recent methods in scribe identification and verification. Scribe identification methods can be divided into two categories: allograph based methods and textual based ones. Although textual based methods are easier to interpret, the best results so far were achieved by allograph based approaches. One such approach is based on GMM supervectors. This method is compared against other allograph based methods on contemporary datasets such as the ICDAR 2013 competition set and the CVL dataset showing TOP-1 accuracy

of more than 97%. Finally, the method has been applied on a set of datum lines of high medieval papal charters. Background artifacts reduce the accuracy of the classification, thus a word based approach built on GMM supervectors, which reduces the error by a large margin, was developed. This also reveals the limit of current datasets which consist of too few scribes and are too clean in contrast to historical documents. However, in general writer identification / verification methods perform very well, especially when they are applied on contemporary documents, and can thus reduce the effort of large-scale identification / verification drastically.

3.3 Algorithmic Paleography

Nachum Dershowitz (Tel Aviv University, IL)

License  Creative Commons BY 3.0 Unported license
© Nachum Dershowitz

Modern algorithms can help in many tasks of interest to scholars of the humanities and, in particular, in the analysis of old manuscripts and texts. We describe ongoing research in the application of methods developed in the fields of computer vision, bioinformatics, and machine learning to endeavors such as the paleographic analysis of manuscripts, finding documents in the same hand, searching within images, and tracing fibers in papyri. Our examples include the Dead Sea Scrolls, the Cairo Genizah, and the Tibetan Buddhist corpus.

3.4 Appearance Modeling for Handwriting Recognition

Gernot Fink (TU Dortmund, DE)

License  Creative Commons BY 3.0 Unported license
© Gernot Fink

Main reference Gernot A. Fink, “Markov Models for Pattern Recognition – From Theory to Applications,” *Advances in Computer Vision and Pattern Recognition*, Springer, London, 2014.

URL <http://www.springer.com/978-1-4471-6307-7>

In this presentation I give an overview of appearance modeling techniques for offline handwriting recognition, i. e., the recognition of handwriting from document images.

I first present the traditional techniques that have been proposed for the recognition of isolated characters and follow a classical pattern recognition pipeline (cf. e. g. [1, Chap. 10–11]).

Then I focus on the recognition of cursive script, where segmentation-based approaches fail due to the very nature of cursive writing and the high variability of the data. Therefore, so-called segmentation-free methods have been proposed, the most well-known being based on hidden Markov models (HMMs) (cf. [2, 5]). I present the general architecture of an HMM-based handwriting recognition system and introduce the sliding-window approach that is essential for converting images of handwritten script into sequences of feature vectors that can be modeled by HMMs. Afterwards, I describe how structured recognition models can be built based on elementary modeling units. For these mostly the characters of the respective script are used but there also exist approaches where context-dependent characters or sub-character units are applied.

In addition to modeling approaches for handwriting recognition I briefly present how script appearance is represented in today’s handwriting retrieval systems that are based

on query-by-example word spotting techniques (cf. [3]). In this field image descriptors based on gradient statistics are used for building holistic models of individual query words following the Bag-of-Features (BoF) principle (cf. [4]). In order to improve the performance beyond basic BoF-based word-spotting systems, the BoF principle can be combined with the sequential statistical modeling provided by HMMs. These BoF HMMs today deliver excellent handwriting retrieval performance [7, 6].

From these considerations it can be concluded that impressive results can be achieved for problems with large, annotated training data sets. Language constraints can be described well statistically, but the training of such models for non-contemporary data remains an open problem. A further challenge is that special attention to character appearance is almost exclusively achieved via preprocessing and feature extraction and there exist no principled approaches for sharing of structural cues between character models. It is especially unclear how to transfer such “appearance knowledge” to different writing styles, from printed to handwritten material, or to an entirely new type of script.

Therefore, from a Pattern Recognition viewpoint it appears to be especially interesting to automatically extract script-specific information from example data, to exploit semi-supervised learning strategies, i. e., to learn appearance models from a few labeled and a huge number of unlabeled samples, and to systematically transfer or adapt appearance models to new tasks. With respect to applications in paleographic research it will be important to involve paleographic experts as humans-in-the-loop such that automatic pattern recognition methods rather provide assistance than try to compute necessarily imperfect finalized solutions.

References

- 1 David Doermann and Karl Tombre, editors. *Handbook of Document Image Processing and Recognition*. Springer, London, 2014.
- 2 Gernot A. Fink. *Markov Models for Pattern Recognition, From Theory to Applications*. Advances in Computer Vision and Pattern Recognition. Springer, London, 2 edition, 2014.
- 3 Josep Lladós, Marçal Rusiñol, Alicia Fornés, David Fernández, and Anjan Dutta. On the influence of word representations for handwritten word spotting in historical documents. *Int. J. Pattern Recognition and Artificial Intelligence*, 26(5), 2012.
- 4 Stephen O’Hara and Bruce A. Draper. Introduction to the bag of features paradigm for image classification and retrieval. *Computing Research Repository*, arXiv:1101.3354v1, 2011.
- 5 Thomas Plötz and Gernot A. Fink. *Markov Models for Handwriting Recognition*. SpringerBriefs in Computer Science. Springer, 2011.
- 6 Leonard Rothacker, Marçal Rusinol, and Gernot A. Fink. Bag-of-features HMMs for segmentation-free word spotting in handwritten documents. In *Proc. Int. Conf. on Document Analysis and Recognition*, Washington DC, USA, 2013.
- 7 Leonard Rothacker, Szilard Vajda, and Gernot A. Fink. Bag-of-features representations for offline handwriting recognition applied to Arabic script. In *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, Bari, Italy, 2012.

3.5 Separating glyphs of handwritings with Diptychon

Björn Gottfried (Universität Bremen, DE)

License  Creative Commons BY 3.0 Unported license
© Björn Gottfried

Joint work of Gottfried, Björn; Lawo, Mathias

My presentation is about a transdisciplinary project in the context of digital palaeography in which methods are developed in order to support palaeographers in comparing handwritings. It is supported by the German Research Foundation, DFG, under grant number GO 2023/4-1 (LA 3066), LA 3007/1-1.

As one important objective the separation of handwritings into their constituent glyphs is discussed and motivated as follows:

- Separated glyphs allow the search for strings in the original document, showing the context of specific glyphs,
- facilitate the character-wise comparison of handwritings, and
- enable the characterisation of the specificities of single glyph images.

Though being generally very difficult and sometimes even impossible, the extraction of single glyphs is challenging but not impossible. An interactive human-machine methodology enables the extraction of single glyphs by combining both the precision and efficiency of the computer as well as the expertise and flexibility of the user. An example of an automatic method is provided in [1].

The methodology has been applied to different handwritings between the 9th and 18th centuries and depends on the specific characteristics of each handwriting. The interaction effort to correct imperfect suggestions provided by the computer lies in the average around 2 seconds per glyph and ranges between 0.6 and 1.4 operations per glyph.

References

- 1 Jan-Hendrik Worch, Mathias Lawo, Björn Gottfried. *Glyph spotting for mediaeval handwritings by template matching*. ACM, Springfield, Paris, France, September 4–7, 2012.

3.6 Deciphering and Mapping the Socio-Cultural Landscape of 12th Century Jerusalem: Texts, Artifacts and Digital Tools

Anna Gutgarts-Weinberger (The Hebrew University of Jerusalem, IL) and Iris Shagrir (The Open University of Israel – Raanana, IL)

License  Creative Commons BY 3.0 Unported license
© Anna Gutgarts-Weinberger and Iris Shagrir

Research on the urban layout of medieval Jerusalem has traditionally been based on the integration of written descriptions and archaeological investigation. Especially privileged in this context were the monumental buildings whose architecture was both described in detail and has been in many cases, still visible on the ground. In the Crusader period in Jerusalem realities changed. First, there was an upsurge in documentation regarding the life in the city. This was produced by various institutions and agents operating in the newly established Christian capital in the Levant. Secondly, we have information not only about the big monuments but also on private buildings, urban zoning, the commercial areas and religious endowments. From this wealth of documents we aim to produce a detailed database which will allow for qualitative and quantitative analysis of the urban configuration and

topographical layout, in a manner that has never been performed before. We aim to use this database to study the social, cultural and perhaps economic development and hopefully clarify further the distribution of various sectors of the population within the city.

Method outline: The project presented here aims at reconstructing and analyzing chronologically and spatially the development of medieval Jerusalem, 11th- 13th centuries, based on an analysis of the entire corpus of legal, historical, descriptive and religious documents pertaining to sites and events in Jerusalem of the crusader period. The plan is to juxtapose textual and archaeological data, derived from excavations conducted over recent decades. To date, no integrative study of medieval Jerusalem exists. The combination of documentary and archaeological data is expected to enable a comprehensive spatial-temporal reconstruction and analysis of the plan, topography, property-ownership and urban development of Jerusalem over this period. The study aims at assimilating up-to-date insights from the Digital Humanities, in order to create an integrative record of spatially positioned archaeological and topographical data captured and represented on a Geographical Information System (GIS), with carefully categorized text-based historical analysis. The project promises to yield results that will greatly augment our understanding of the history of the Holy City, and generate new questions and further research. Considering the different nature and number of the available sources, the main challenge in the construction of the database lies in the conversion and standardization of historical and archaeological sources into data that can be collated and analyzed from a chronological as well as spatial perspective. This can be demonstrated on the documents pertaining to the city during the period in question. These documents record transactions involving exchanges of properties in and around the city of Jerusalem, conducted among various agents. In order to isolate and trace multiple strands of information, the documents were collected and organized according to their chronological order and the geographic information they hold. They were then broken down into multiple subcategories according to several main thematic clusters, among which are agency, institutional association, property details and connections to other documents. This deconstruction of the documents into their primary elements is designed to accommodate for multifaceted cross-sectioning of the data, allowing an examination and analysis of correlations between multiple clusters of information, thus incorporating both chronological and spatial evolution. This type of analysis yields a detailed and dynamic representation of the underlying mechanisms responsible for the changes that occurred in the cityscape throughout the 12th century. It also reflects the balance and relationship between socio-economic functions and the urban setting they inhabited, helping deciphering and better understanding Frankish Jerusalem's urban fabric.

Sample issues/challenges for DH: Developing software tools that support the process of interpretation and digital tools to complement the human expertise in actions such as:

- Cross-referencing narrative and archeological data.
- Representation of static vs. dynamic data.
- Representation of discrete objects vs. abstractions.
- Codifying and calibrating non-specific property descriptions.
- Automatic identification of different name variants.
- Isolation, classification and analysis of transactions, and statistical significance.

3.7 Positioning computational tools

Tal Hassner (The Open University of Israel – Raanana, IL)

License © Creative Commons BY 3.0 Unported license
© Tal Hassner

Main reference T. Hassner, L. Wolf, N. Dershowitz, “OCR-free Transcript Alignment,” in Proc. of the 12th Int’l Conf. on Document Analysis and Recognition (ICDAR’13), pp. 1310–1314, IEEE, 2013; pre-print available from author’s webpage.

URL <http://dx.doi.org/10.1109/ICDAR.2013.265>

URL http://www.openu.ac.il/home/hassner/projects/Ofta/ofta_online.pdf

URL <http://www.openu.ac.il/home/hassner/projects/Ofta/>

The conclusions of the Schloss Dagstuhl – Leibniz Center for Informatics, Perspective Workshop on “Computation and Palaeography: Potentials and Limits”, 2012, expressed in its subsequent manifesto [1], listed a number of crucial points of concern regarding the collaboration between computer scientists and palaeographers. In my talk, I focus on two of these, namely data availability and its significance to the development and training of computerized systems; and the so-called “black-box” issue, relating to the need of palaeography scholars to have more understanding and interaction with their computerized tools. Taking as an example the specific task of transcript alignment, I attempt to draw a taxonomy of available computerized tools, based on the data required to train them versus the amount of interaction they require of the scholar. The key question raised is where in this taxonomy would an ideal computerized palaeographic tool be positioned, in order for it to be both realistic in its prerequisite data and effective in its capabilities?

As a potential answer, I provide the recently developed OCR-Free transcript alignment system [2]. This system directly matches the pixels in an image of a historical text with those of a synthetic image created from the transcript for the purpose. This, rather than attempting to recognize individual letters in the manuscript image using optical character recognition (OCR). It therefore does not require manual labeling or pre-segmentation of letters nor massive training data required to learn particular alphabets and characteristics of scribal hands. I visualize the output of this system and discuss the ways in which it may be manipulated by the scholar in order to quickly and effectively correct for alignment errors. I conclude with suggesting future work, discussing how such corrections can potentially be used to learn, on the fly, the particular characteristics of the manuscript at hand, and improve alignment from one line of text to the next.

References

- 1 Hassner, T., Rehbein, M., Stokes, P.A., Wolf, L.: Computation and Palaeography: Potentials and Limits (Dagstuhl Perspectives Workshop 12382). Dagstuhl Manifestos **2** (2013)
- 2 Hassner, T., Wolf, L., Dershowitz, N.: OCR-free transcript alignment. In: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, IEEE (2013), pp. 1310–1314

3.8 DIVADIA & HisDoc 2.0 Approaches at the University of Fribourg to Digital Paleography

Marcus Liwicki (DFKI – Kaiserslautern, DE)

License  Creative Commons BY 3.0 Unported license
© Marcus Liwicki

Joint work of Liwicki, Marcus; Garz, Angelika; Ingold, Rolf; Wei, Hao; Chen, Kai; Eichenberger, Nicole

In this article we present DIVADIA, a toolkit for labeling medieval documents and the HisDoc and HisDoc 2.0 projects on Document Image Analysis (DIA) funded by the Swiss National Science Foundation (SNSF). At the University of Fribourg, we conceptualize a workspace comprising methods for input and presentation of humanists' research on historical documents. The underlying architecture of the workspace consists of three modules concerned with Item Description, Content Representation, and Research Data. Each of the modules provides computational methods for semi-automatic processing of document images, transcriptions, annotations, and research data. DIVADIA is ongoing research at DIVA research group at the University of Fribourg. In its current state it provides Document Image Analysis (DIA) methods for layout analysis, script analysis, and text recognition of historical documents. The methods build on the concept of incremental learning and provide users with semi-automatic labeling of document parts, such as text, images, and initials. The future goal is to provide means for labelling, annotating, searching, browsing, viewing, and comparing documents as well as presenting research data in adequate visualizations. In the HisDoc projects we perform research on textual heritage preservation. HisDoc aimed at layout and textual content analysis of historical documents, i. e., focusing on philological studies. HisDoc 2.0 will take the approach a step further: it will be dedicated to paleographical studies and incorporate semantic domain knowledge automatically extracted from existing document databases into DIA methods in order to facilitate large-scale processing. As such we will investigate the yet missing ingredients for automatic large-scale analysis of historical documents, and how to make the results useful for historians. While concentrating on medieval manuscripts, we intend to develop methods easily adaptable to other kinds of documents and scripts. During the discussion we presented the current stage of the DIVA-HisDB which will contain larger amounts of annotated historical images with difficult layouts. Every year during the ICDAR and ICFHR conferences we will publish new data along with a benchmark competition. An interesting discussion point raised during the seminar was the presentation of document processing results; as developer of document enhancement methods we should make it clear that the output of the enhancement method (e. g., binarization) is a processed image and not a direct photograph of the original document. The main reason for that is that each data processing step introduces derivations from the original image and might also introduce errors. In the worst case a paleographer investigating only a processed image without being aware of the processing steps might draw conclusions which would not have been drawn when investigating the original physical document.

3.9 Word spotting in historical manuscripts. The “Five Centuries of Marriages” project

Josep Lladós (Autonomous University of Barcelona, ES)

License © Creative Commons BY 3.0 Unported license
© Josep Lladós

Search centered at people is very important in historical research, including historical demography, people trajectories reconstruction and genealogical research. Queries about a person and his/her connections to other people allow to get a picture of a historical context: a person’s life, an event, a location at some period of time. For this purpose, scholars use documents like birth, marriage, or census records.

From a technical point of view, word spotting plays a central role in searching among historical people records. Word spotting is the process of retrieving all instances of a queried keyword from a digital library of document images. We have proposed different word spotting approaches for historical manuscript retrieval. In particular, we have evaluated the performance within the EU-ERC project Five Centuries of Marriages (5CofM), which consists in the analysis of marriage license records from the Barcelona Cathedral.

We have made some contributions in context-aware word spotting. Usually word spotting is built based solely on the statistics of local terms. The use of correlative semantic labels between codewords adds more discriminability in the process. Three levels of context can be defined in a word spotting scenario. First, the joint occurrence of words in a given image segment. Second, the geometric context involving a language model regarding to the relative 1D or 2D position of objects. Third, the semantic context defined by the topic of the document. A number of document collections convey an underlying structure.

We take advantage of the structure to boost the search of words, with a joint search of the query word and its context.

3.10 Modern Technologies for Manuscript Research

Robert Sablatnig (TU Wien, AT)

License © Creative Commons BY 3.0 Unported license
© Robert Sablatnig

Joint work of Miklas, Heinz; Schreiner, Manfred; Čamba, Ana; Hürner, Dana; Vetter, Willi; Garz, Angelika; Sablatnig, Robert

Main reference S. Fiel, R. Sablatnig, “Writer Identification and Writer Retrieval Using the Fisher Vector on Visual Vocabularies,” in Proc. of the 12 Int’l Conf. on Document Analysis and Recognition (ICDAR’13), pp. 545–549, IEEE, 2013.

URL <http://dx.doi.org/10.1109/ICDAR.2013.114>

URL <http://caa.tuwien.ac.at/cvl/research/sinai/index.html>

Manuscript analysis and reconstruction has long been solely the domain of philologists who had to cope with complex tasks without the aid of specialized tools. Technical scientists were only engaged in recording and conservation of valuable objects. In recent years, however, interdisciplinary work has constantly gained importance, concentrating not on a few special tasks only, like the development of OCR software, but comprising an increasing amount of relevant interdisciplinary fields like material analysis and document reconstruction. It may be expected that in the long run the decipherment, study and edition of such sources will predominantly be done based on digital images. This relieves the originals, makes their investigation independent of the place of preservation and permits a lossless storage of the

contents. Additionally, a more precise and less time-consuming investigation of manuscripts through automatic image analysis is made possible. Especially for information invisible to the human eye, spectral imaging methods are applied in order to visualize lost content. Digital cameras sensitive to an extended spectral band are used to produce multi-spectral images which (in combination with digital image processing) allow enhancing the readability of “hidden” texts and an automated investigation of structure and content of the manuscripts.

In order to acquire manuscripts in libraries, a system is needed that is easily portable, robust, and permits quick handling and fast imaging. Thus, we combined a Nikon D2Xs RGB camera to obtain conventional color images and a Hamamatsu C9300-124 high resolution camera with a spectral response from Ultra-Violet (UV) to Near-Infra-Red (NIR, 330 to 1000 nm) and a resolution of 4000x2672 pixels. The lighting system consists of two LED panels with 13 narrow spectral bands. Additionally, four white light LED panels are used for the RGB photographs, since LED lighting does not impose additional heat radiation on the manuscript.

A multi-spectral representation of the page (one object in multiple spectral ranges) acquired in this manner is the basis for our subsequent analyses like image enhancement, since this data representation holds a great potential for increasing the readability of historic texts, especially if the manuscripts are (partially) damaged and consequently hard to read. The readability enhancement is based on a combination of spatial and spectral information of the multivariate image data, a so called Multivariate Spatial Correlation (MSC). The benefit of this method is the possibility to specifically consider individual text regions in document images. Additionally, Independent Component Analysis (ICA), Principal Component Analysis (PCA), and Fisher Linear Discriminate Analysis (LDA) have been successfully applied in order to reduce the dimension of the multispectral scan and for the separation and enhancement of diverse writings. Since LDA is a supervised dimension reduction tool, it is necessary to label a subset of multispectral data. For this purpose, a semi-automated label generation step was developed, which is based on an automated detection of text lines. Thus, the approach is not only based on spectral information – like PCA and ICA – but also on spatial information. A qualitative analysis shows, that the LDA based dimension reduction gains better performance, compared to unsupervised techniques.

Another interesting aspect when working with manuscripts is the automatic identification of authors based on their scribes. We investigated scribe identification on the example of historical Slavonic manuscripts. The quality of these documents is partially degraded by faded-out ink or varying background. The writer identification method used is based on textual features, which are described with Scale Invariant Feature Transform (SIFT) features. A visual vocabulary is used for the description of handwriting characteristics, whereby the features are clustered using a Gaussian Mixture Model and employing the Fisher kernel. The writer identification approach is originally designed for grayscale images of modern handwritings. But contrary to modern documents, the historical manuscripts are partially corrupted by background clutter and water stains. As a result, SIFT features are also found on the background. Since the method shows also good results on binarized images of modern handwritings, the approach was additionally applied on binarized images of the ancient writings. Experiments show that this preprocessing step leads to a significant performance increase: The identification rate on binarized images is 98.9%, compared to an identification rate of 87.6% gained on grayscale images.

References

- 1 Fabian Hollaus and Melanie Gau and Robert Sablatnig,, *Enhancement of Multispectral Images of Degraded Documents by Employing Spatial Information*. Proc. of 12th International

- Conference on Document Analysis and Recognition (ICDAR 2013). 2013, pp. 145–149
- 2 Fabian Hollaus and Melanie Gau and Robert Sablatnig, *Acquisition and Enhancement of Multispectral Images of Ancient Manuscripts*, Proc. of 11th Culture and Computer Science Conference, 2013, ed. Sieck, J., Franken-Wendelstorf, R.

3.11 tranScriptorium

Joan Andreu Sanchez Peiro (Polytechnic University of Valencia, ES)

License © Creative Commons BY 3.0 Unported license
© Joan Andreu Sanchez Peiro

Joint work of J. A. Sanchez Peiro; G. Mühlberger; B. Gatos; P. Schofield; K. Depuydt; R. M. Davis; E. Vidal; J. de Does

Main reference J. A. Sanchez Peiro, G. Mühlberger, B. Gatos, P. Schofield, K. Depuydt, R. M. Davis, E. Vidal, J. de Does, “tranScriptorium: a european project on handwritten text recognition,” in Proc. of the 2013 ACM Symp. on Document Engineering, pp. 227–228, ACM, 2013.

URL <http://dx.doi.org/10.1145/2494266.2494294>

tranScriptorium (<http://www.transcriptorium.eu>) [1] aims to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using modern, holistic Handwritten Text Recognition (HTR) technology.

tranScriptorium will turn HTR technology into a mature technology by addressing the following objectives:

1. Enhancing HTR technology for efficient transcription.
Departing from state-of-the-art HTR approaches, tranScriptorium will capitalize on interactive-predictive techniques for effective and user-friendly computer-assisted transcription.
2. Bringing the HTR technology to users.
Expected users of the HTR technology belong mainly to two groups: a) individual researchers with experience in handwritten documents transcription interested in transcribing specific documents. b) volunteers which collaborate in large transcription projects.
3. Integrating the HTR results in public web portals.
The HTR technology will become a support in the digitization of the handwritten materials. The outcomes of the tranScriptorium tools will be attached to the published handwritten document images. This includes not only full, correct transcriptions, but also partially correct transcription and other kinds of automatically produced metadata, useful for indexing and searching.

References

- 1 J. A. Sanchez and G. Mühlberger and B. Gatos and P. Schofield and K. Depuydt, R. M. Davis and E. Vidal and J. de Does. *tranScriptorium: a European Project on Handwritten Text Recognition*. ACM Symp. on Document Engineering DOCENG, 2013, pp. 227–228.

3.12 Text Classification and Medieval Literary Genres

Wendy Scase (University of Birmingham, GB)

License © Creative Commons BY 3.0 Unported license
© Wendy Scase

This presentation reported on investigation of problems in text classification that are experienced in the creation and querying of large corpora of texts and images. Humanists know that genre information is relevant to the palaeographical analysis of documents. For

example, the genre of a text can influence the scribe's choice of script. A legal document may be written in a cursive script, reflecting the need for speed and economy in the production of the document, whereas a bible will often be written in a formal script that requires the scribe to create letters from many small, careful strokes. This choice may reflect the aspirations of the patron (to save his soul, or display his wealth), the need for the scribe to conceal his identity (where the manuscript may be considered heretical) and so on. So classification by genre is relevant to the interpretation of material in corpora, especially where the corpora are produced from many different parent resources (e.g. archives of documents, collections of literary texts, chronicles). Classification of genres from the medieval point of view is however still little understood. A further problem occurs when resources from different modern genres are federated in a resource (e.g. dictionaries, catalogues, full-text transcriptions). The user needs to know the genre of the text retrieved to interpret it accurately. Manuscripts Online (www.manuscriptsonline.org) is an experiment with federating resources relating to medieval; British texts was used to illustrate these problems and some partial solutions. More work needs to be done. The final part of the presentation reported on work towards the expansion and further enhancement of a corpus reported on at Dagstuhl Perspectives Workshop 12382. The Vernon manuscript scribe's text and image corpus (Bodleian Library, MS Eng. Poet.a.1) has been increased with the digitisation of the Simeon manuscript (British Library, Addit. MS 22283), also partly copied by the Vernon scribe. Many research questions could be explored if the images could be provided with aligned transcription. The presentation proposed that the existing files of the Vernon manuscript project could be harnessed to create a training set that would permit semi-automated labelling of the images of the Simeon manuscript.

3.13 Describing Handwriting – Again

Peter A. Stokes (King's College – London, GB)

License © Creative Commons BY 3.0 Unported license
© Peter A. Stokes

Joint work of Stokes, Peter A.; Brookes, Stewart; Noël, Geoffroy; Buomprisco, Giancarlo; Watson, Matilda; Matos, Debora

Main reference DigiPal: Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic. London: 2011–14.

URL <http://www.digipal.eu/>

When considering the identification of characters and scripts, two important aspects that were identified in the 2012 Dagstuhl Perspectives Workshop on computing and palaeography are ontologies and mid-level features [1]. This paper focussed on those two aspects, partly in deliberate contrast to the highly computational approach that most studies in the field have taken to date.

To this end, problems not only of terminology but also of conceptual ambiguity and imprecision in palaeography were introduced. The ontology developed for the DigiPal project was briefly presented as a response to this, including the way that it has been used in practice for describing writing in the Latin and Hebrew alphabets as well as for decoration [2, 3]; initial work has also been done to use it for Greek and Latin inscriptions and cursive Latin script. The ontology was presented not as an ideal solution but rather as a pragmatic one that has proven useful in a variety of circumstances, and as a starting-point to a very difficult problem with many challenges that still remain.

The second part of the talk considered possible mid-level features, presenting a selection of potential characteristics of handwriting that are relevant to palaeographers and that seem to this author to be relatively easily amenable to computational analysis but which seem not

to have been considered in practice. These included ‘stabbing’ strokes (perhaps indicating a scribe accustomed to writing on wax), ‘equilibrium’ (the regularity or otherwise of strokes, perhaps a sign of fluency, experience, forgery or imitation), and the effective visualisation of these particularly in the context of other factors such as the codicological structure of the book. As an aside, DigiPal’s RESTful API was also introduced as a potential source of annotated images for the training of computer vision systems.

None of these methods or approaches is necessarily appropriate for writer identification, but they suggest other directions in which computer vision might be taken and which perhaps are more pertinent to research in medieval manuscripts than some of the work done to date.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7) under grant agreement no. 263751.

References

- 1 T. Hassner, M. Rehbein, P. A. Stokes, L. Wolf (eds). Computation and Palaeography: Potentials and Limits. *Dagstuhl Manifestos* 2(1):14–35, 2013. DOI: 10.4230/DagMan.2.1.14
- 2 *DigiPal: Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic*. London: 2011–14. <http://www.digipal.eu/>
- 3 P. A. Stokes, S. Brookes, G. Noël, G. Buomprisco, D. Matos and M. Watson. The DigiPal Framework for Script and Image. *Digital Humanities 2014 Book of Abstracts* (Lausanne, 2014), pp. 541–3. <http://dharchive.org/paper/DH2014/Poster-193.xml>

3.14 Bridging the gap between Digital Palaeography and Computational Humanities

Dominique Stutzmann (Institut de Recherche et d’Histoire des Textes (CNRS) – Paris, FR)

License  Creative Commons BY 3.0 Unported license
© Dominique Stutzmann

As part of the common introduction to seminars 14301 (Computational Humanities – bridging the gap between Computer Science and Digital Humanities) and 14302 (Digital Palaeography New Machines and Old Texts), the first paper presented the specific field of the Digital Humanities devoted to the history of scripts, aka “digital palaeography” and why it is of interest even for textual scholars. Texts are transmitted through signs; signs are transmitted through shapes; the shapes for each sign evolve and are perceived for their meaning and in their historical context. Moreover, scripts convey a particular meaning for themselves, as do the *litterae elongatae* and diplomatic script in a diploma of Charlemagne, referring to imperial *litterae caelestes* and supporting the claim of a new Empire, while the same emperor, on the other hand, could support the Caroline script, named after him. Issues for the palaeographer encompass the history of script, cultural history, writer identification, dating and assigning a place of origin for any written sample. As demonstrated by the examples from the transmission of Cicero’s works, textual scholarship need to envision the materiality of the transmitted text (not least for classical texts for which there are only medieval witnesses) and digital palaeography addresses the notions of text through image, layout and shape, through their materiality, their history, origin and provenance of the witnesses, through their cultural significance. Digital Palaeography means: how to use computers to help the humanities identifying the relevant historical phenomena, to identify interscript, interscribal, intra-script and intra-scribal variations as well as cultural and textual relevant features. Some bridges with Computational Humanities are obvious: Keyword Spotting and retrieval is similar to indexing techniques; Handwritten Text Recognition is

linked to scholarly editing textual transmission, ideas and their reception. In the issues raised by 14301 are mentioned the kind of results and the transfer to other fields (methodology and applicability), the difficulties in cross-disciplinary collaboration, the human resources and communication, the variability and quality of data, the evaluation and ground-truth. Demonstration and proof in the Humanities and Computer Science, or the measure of success supposes a unique ground-truth, which does not always exist, while the result of a calculation generally represents only an additional clue in the complex reality. All these issues, as well as the crucial notion of reciprocal uncertainties have been addressed in the perspective workshop 12832. Indeed, the four core issues identified issues in 2012 were “Communication and roles in the interdisciplinary interplay, the notions of black box and meaning of calculation”, the evaluation of and need for “quality and quantity” in the data from the humanities, and the new audiences (with correlations in interoperability, rights managements and engaging with other communities). These issues are now to be addressed by Digital Palaeographers on a technical and epistemological level, but are also common to all fields in the Digital Humanities and should appeal for a more intense dialogue.

3.15 Digital Palaeography. Text-Image Alignment and Script/Scribal Variability (ANR ORIFLAMMS / Cap Digital)

Dominique Stutzmann (Institut de Recherche et d'Histoire des Textes (CNRS) – Paris, FR)

License © Creative Commons BY 3.0 Unported license
© Dominique Stutzmann

Joint work of Stutzmann, Dominique; Lavrentiev, Alexei; Kermorvant, Christopher; Bluche, Théodore; Leydier, Yann; Ceccherini, Irene; Eglin, Véronique; Vincent, Nicole; Debiais, Vincent; Treffort, Cécile; Ingrand-Varenne, Estelle; Smith, Marc

URL <http://oriflamms.hypotheses.org/>

URL [http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2\[CODE\]=ANR-12-CORP-0010](http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-CORP-0010)

Medieval scripts are a challenge to historical analysis, as for describing and representing the graphical evidence, analyzing and clustering letter forms and their features through Computer Vision and analyzing historical phenomena. The ANR funded research project ORIFLAMMS (Ontology Research, Image Feature, Letterform Analysis on Multilingual Medieval Scripts, 2013-2016) gathers seven partners from the Humanities and Computer Science (IRHT = Institut de Recherche et d'Histoire des Textes, CNRS; CESCUM = Centre d'Études Supérieures de Civilisation Médiévale; École Nationale des Chartes; ICAR = Interactions Corpus Apprentissages Représentations, École Normale Supérieure de Lyon, for the Humanities; A2iA; LIRIS = Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSA Lyon; LIPADE = Laboratoire d'Informatique de Paris Descartes, for Computer Science). It aims at studying the coherence and variability of graphical systems, according to their language, level of formality, support, genre, date and place, as well as creating an ontology of medieval signs, through the alignment of text and images, by extracting letterforms, abbreviations and signs, then perform pattern similarity analysis, and enhance the results with computational linguistics and paleographical analysis. In order to achieve representative results, several core corpuses have been identified (charters, books, books of charters such as cartularies and registers, inscriptions). The research is based on XML-TEI compliant editions and compels to deepening our understanding of scribal systems and forms [1, 2]. As part of this research, a software has been developed in order to easily visualize and validate the text-image alignment. The latter is produced by two different systems developed in this project: the first one without prior knowledge [3], the second one

with GMM and DNN with very good results. By now, two large data sets have been aligned: *Queste du Graal* including 130 pages, 10700 lines, more than 115'000 words and 400'300 characters; *Fontenay* including 104 pages, 1341 lines, more than 22'200 words and 99'900 characters. This is a major first step. With the following corpuses, this research contributes to both Humanities (letterform identification, historical semiotics) and Computer Science (Handwriting recognition), with the core idea of not reinventing the wheel, but using former research, computer and human brain at their maximal capacities.

Acknowledgements. The research leading to these results has received funding from the Agence Nationale de la Recherche and Cap Digital under grant agreement no. ANR-12-CORP-0010.

References

- 1 D. Stutzmann. Paléographie statistique pour décrire, identifier, dater . . . Normaliser pour coopérer et aller plus loin? In *Kodikologie und Paläographie im digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*, Norderstedt, 2010, pp. 247–277
- 2 D. Stutzmann. Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS, In *Document numérique*, 16/3 (2013):81–95. DOI: 10.3166/DN.16.3.69-79.
- 3 Y. Leydier, V. Eglin, S. Bres, D. Stutzmann. Learning-free text-image alignment for medieval manuscripts. In *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, Crete, Greece, 2014.

3.16 Digital Images of Ancient Textual Artefacts: Connecting Computational Processing and Cognitive Processes

Ségolène Tarte (*University of Oxford, GB*)

License © Creative Commons BY 3.0 Unported license
© Ségolène Tarte

Main reference S. Tarte, “Interpreting Textual Artefacts: Cognitive Insights into Expert Practices,” in *Proc. of the 2012 Digital Humanities Congress*, 2012.

URL <http://www.hrionline.ac.uk/openbook/chapter/dhc2012-tarte>

Drawing on examples from palaeographical scholarship rooted in Classics and in Assyriology, this talk will give an overview of how it might be possible to connect computational processing and cognitive processes. As a preamble, considering the type of material that palaeographers (be they Classicists, Mediaevalists, or Assyriologists) work from, I will argue that an image of an ancient textual artefact is a digital avatar of the textual artefact. In digital palaeography, these images are an absolute prerequisite, but it is crucial to be aware that as avatars they are already part of the interpretative workflow that transforms the data (the textual artefact) into knowledge and meaning. Digital avatars are interpretative; they express a certain form of presence of the textual artefact, they are contingent on the act of digitization and they have an expected performative value [1]. All those implicit aspects that participate in the act of knowledge creation coexist with the intuitive strategies that scholars develop to carry out their task. I will present three such strategies identified through ethnographic studies of Classicists and Assyriologists at work [2]. Establishing a correspondence between these ethnographic observations and cognitive processes (as identified in the cognitive sciences literature), I will show examples of how these cognitive processes influenced and supported the choice of computational processing made by the scholars. Namely: embodied cognition and an awareness of the materiality of a papyrus suggested modelling it as a roll to justify the repositioning of a fragment; kinaesthetic facilitation was supported through digital tracing of the text of another artefact, thereby supporting the establishment of the connection

between the text as a shape and the text as a meaning; depth perception through monocular parallax motion was supported for yet another artefact by the digitization process, allow to interactively relight the artefact. These examples are vivid illustrations of the fact that understanding scholars' cognitive involvement have the exciting potential to facilitate the seamless integration of the use of computational tools within the research workflow whilst at the same time supporting embodied sense-making practices.

References

- 1 Tarte, Ségolène M. The Digital Existence of Words and Pictures: The Case of the Artemidorus Papyrus In *Historia 3:61*, pp. 325–336 (+bibliog. pp. 357-61; fig. pp 363-5), 2012.
- 2 Tarte, Ségolène, Interpreting Textual Artefacts: Cognitive Insights into Expert Practices In: *Proc. of the Digital Humanities Congress 2012*, Ed. Clare Mills, Michael Pidd, and Esther Ward, Sheffield: HRI Online Publications, Studies in the Digital Humanities, 2014.

3.17 Text classification

Nicole Vincent (Paris Descartes University, FR)

License © Creative Commons BY 3.0 Unported license
© Nicole Vincent

Classification, and text classification, has to be done with respect to some objectives. These objectives are varying according to the field of interest possibly being medical, security or palaeography. Some questions are rising, such as: Do you have some ground truth available defining the classes and their number? One point is the definition of features. But how to choose them? Choose many to have a large amount of information. Not too many because of dimensionality problem and because the aim is to decrease complexity. What about feature selection? What about learning? What may be the criteria to choose features: have they to be understandable? Should they be local or global, addressing details? Should they be invariant towards different factors? What about the process? Defined by the expert, blind based on computer science theory, based on pixels or features or primitives, involving an interaction with the user? 4 examples of text classification are presented. They have been developed in the GRAPHEM project funded by French National Research Agency:

- One involving a decomposition of writing that models the way the drawing is done
- One based on the statistical analysis of the writing contour
- One trying to be the automated version of an expert palaeographer
- One base on the statistical analysis of some low level patterns

3.18 Diplomatics and Digital Palaeography

Georg Vogeler (*Karl-Franzens-Universität Graz, AT*)

License © Creative Commons BY 3.0 Unported license
© Georg Vogeler

Main reference A. Ambrosio, S. Barret, G. Vogeler (eds.), “Digital diplomatics. The computer as a tool for the diplomatist?” Wien, Böhlau, 2014.

URL <http://www.boehlau-verlag.com/978-3-412-22280-2.html>

Manuscripts documenting a single legal act usually authenticated by special means are an important source for the history of the middle ages and the early modern period. They are subject of the research field of “diplomatics”, which includes skills in philology, sphragistics, chronology and certainly palaeography. The paper gave an overview on the issues of image based digital methods in diplomatics and their applicability to digital palaeography, and addressed the following questions: what *diplomatics* can contribute to *Digital Palaeography*? Should/Can we build an integrated “Virtual Research Environment” for digital palaeography?

- Digital Palaeography applied to diplomatic sources confronts new challenges in comparison with literary manuscripts, since charters are short, very numerous, formulaic. However, there is usually substantial context information and metadata (date, place).
- Diplomatic writings document the history of Latin script in a specific manner (multiple hands / multiple scripts on one document; “documentary writing style”, functional scripts, chancery scripts vs. notarial hands, stylistic influences between book and diplomatic scripts)
- Large digital charter collections and diplomatic databases like <http://monasterium.net/> offer new possibilities for research in the field of digital palaeography (discovering imitations, forgeries, copies; identifying “writing landscapes”)
- The recently started project “Illuminated Charters” (<http://illuminierte-urkunden.uni-graz.at>) demonstrates how legal instruments may be considered by their value for art history. It allowed to discuss the basic functionalities for an VRE to be used in the project and the role of controlled vocabularies/formal ontologies in these contexts.

The paper demonstrated that legal documents (“charters”, “legal instruments”) are a rich source for experiments with digital methods on historical sources as they convey a large data set with relatively precise historical metadata (date, place, partially even writer) and suggested to work on the definition of interfaces and standards to reuse software tools in a web based palaeographic tool chain, also in order to build “trust” under a cognitive aspect (how does the tool shape the perception of the task?).

4 A Graphical Representation of the Discussed Subjects

The mind map in Fig. 1 presents an overview of the subjects that were broached during the seminar. Each item and sub-item represents an area in which substantial efforts might be concentrated in the future to further research in computational palaeography.

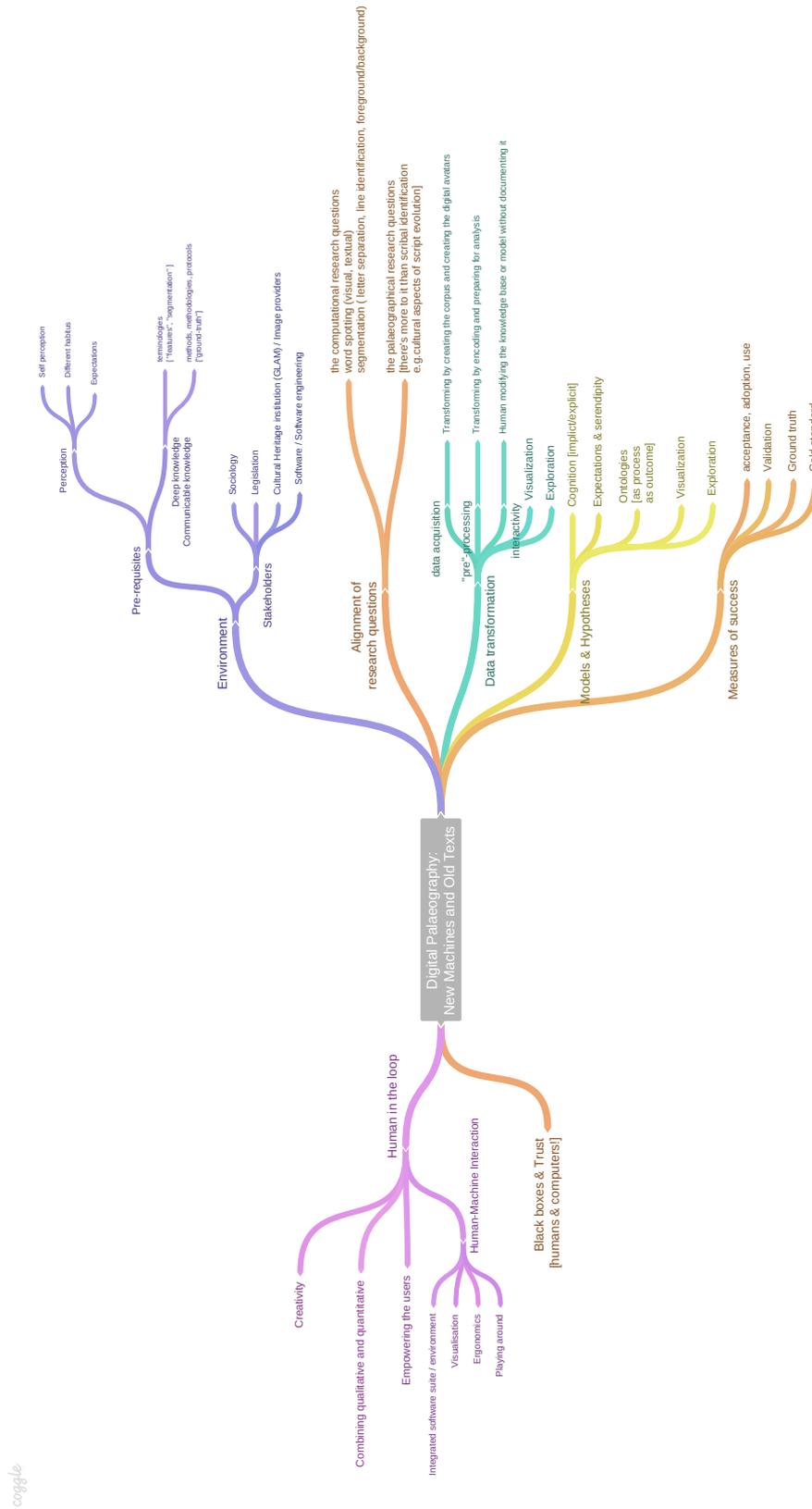


Figure 1 Overview of the themes and issues discussed during the seminar.

Participants

- Orna Almogi
Universität Hamburg, DE
- Vincent Christlein
Univ. Erlangen-Nürnberg, DE
- Nachum Dershowitz
Tel Aviv University, IL
- Véronique Eglin
INRIA / INSA – Lyon, FR
- Jihad El-Sana
Ben Gurion University –
Beer Sheva, IL
- Gernot Fink
TU Dortmund, DE
- Björn Gottfried
Universität Bremen, DE
- Anna Gutgarts-Weinberger
The Hebrew University of
Jerusalem, IL
- Tal Hassner
The Open University of Israel –
Raanana, IL
- Rolf Ingold
University of Fribourg, CH
- Noga Levy
Tel Aviv University, IL
- Marcus Liwicki
DFKI – Kaiserslautern, DE
- Josep Lladós
Autonomous University of
Barcelona, ES
- Frederike Neuber
Karl-Franzens-Univ. Graz, AT
- Jean-Marc Ogier
University of La Rochelle, FR
- Robert Sablatnig
TU Wien, AT
- Joan Andreu Sanchez Peiro
Polytechnic University of
Valencia, ES
- Wendy Scase
University of Birmingham, GB
- Iris Shagrir
The Open University of Israel –
Raanana, IL
- Peter A. Stokes
King's College – London, GB
- Dominique Stutzmann
Institut de Recherche et
d'Histoire des Textes (CNRS) –
Paris, FR
- Ségolène Tarte
University of Oxford, GB
- Nicole Vincent
Paris Descartes University, FR
- Georg Vogeler
Karl-Franzens-Univ. Graz, AT

