DAGSTUHL
REPORTS

Volume 2, Issue 9, September 2012

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/dagrep

Publication date
February, 2013

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `http://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.2.9.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

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

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

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 12361

Information-centric networking – Ready for the real world?

Edited by

Ali Ghodsi[1], Börje Ohlman[2], Jörg Ott[3], Ignacio Solis[3], and Matthias Wählisch[5]

1 **University of California – Berkeley, US**
2 **Ericsson Research – Stockholm, SE,** borje.ohlman@ericsson.com
3 **Aalto University, FI**
3 **PARC – Palo Alto, US,** Ignacio.Solis@parc.com
5 **Freie Universität Berlin, DE,** waehlisch@ieee.org

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 12361 "Information-centric networking – Ready for the real world?". The outcome of this seminar is based on individual talks, group work, and significant discussions among all participants. The topics range from application and performance aspects up to business, legal, and deployment questions. Even though significant progress is visible from the last Dagstuhl Seminar about ICN, there are still thrilling open research questions in all topic areas.

Seminar 02.–09. August, 2012 – www.dagstuhl.de/12361
1998 ACM Subject Classification C.2.1 Network Architecture and Design, C.2.5 Local and Wide-Area Networks—Internet
Keywords and phrases Information-centric, Network architecture, Application structure, Internet business models
Digital Object Identifier 10.4230/DagRep.2.9.1

1 Executive Summary

Ali Ghodsi
Börje Ohlman
Jörg Ott
Ignacio Solis
Matthias Wählisch

Information-centric networking (ICN) defines a communication paradigm that recognizes the dominant usage of the Internet as a substrate to disseminate and access content of all kinds: from traditional web pages to online social networks to file distribution to live and on-demand media feeds. With ICN, the focus shifts from the location at which a content object is stored (typically some server) to the object itself so that scale, efficiency, and robustness of content publication and retrieval can be improved beyond what current Content Distribution Networks (CDNs) can deliver.

Diverse instances of ICN networking architectures were developed, including CCN/NDN, NetInf, DONA, and LIPSIN, among others, and see experimentation at different scale in both academia and industry. The fundamental concepts of ICN have gained popularity in the

Information-centric networking – Ready for the real world?, *Dagstuhl Reports*, Vol. 2, Issue 9, pp. 1–14
Editors: Ali Ghodsi, Börje Ohlman, and Ignacio Solis
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

research community and have been taken up by several research activities that are addressing the topic from different angles.

Numerous research problems remain open, some of which (such as naming content) may find different (optimal) solutions in different deployments while others are more fundamental in nature and could affect the performance of all deployments. The latter include the performance benefits achievable through (cooperative) caching and caching at different points in the network, parallel content retrieval from multiple sources, and tradeoffs between native network layer and overlay-based ICNs. This second Dagstuhl Seminar on information-centric networking is intended to operate as a catalyst for these activities and provide a forum for discussing a selected subset of important research topics that have been identified so far. It will bring together researchers from different ICN backgrounds to discuss fundamentals that matter across the various platforms with the meta goal of identifying obstacles to be overcome, solutions, and paths towards real-world deployments.

In this seminar, we discussed the following core topics: (1) ICN applications and services, (2) ICN performance and comparison of alternative technologies, (3) business, legal, and deployment aspects.

2 Table of Contents

3 Overview of Talks

3.1 SAIL NetInf global connectivity routing and forwarding

Bengt Ahlgren (SICS, SE)

The SAIL NetInf approach to global routing is a hybrid scheme using name-based routing assisted by routing hints. There are two levels of aggregation of routing information in order to make the system scalable, and to improve forwarding performance. First named data objects are grouped into aggregates using the authority part of the name. These aggregates are then associated with a set of routing hints through a lookup service. Requests for a named data object are forwarded using the highest priority hint in the set a router has a forwarding entry for leading towards a network location where the publisher makes the object available. The hint priority implements the second level of aggregation, as only the lowest priority hints need to be announced in the global routing system.

3.2 ICN from a Service Provider Perspective

Marcus Brunner (Swisscom AG – Bern, CH)

The slides are assessing the ICN approach from a service provider or operator perspective. This is done from a technical as well as a first try on the business aspects of an operator. Thought the system designs are different and not finally specified this is just a first approximation of some of the operator thinking behind the abstract approach of ICN. Many of the assumptions made are today's assumptions and might not be true anymore in the future depending on the future development of operator infrastructure.

3.3 Testbeds and mounting large-scale demonstrations: Experiences with NDN

Patrick Crowley (Washington University, US)

Joint work of NDN team

In this talk, we will share both our motivations and methods for conducting large-scale demonstrations of the NDN architecture. We will also introduce open-source tools that we believe will enable others to mount similar demonstrations.

3.4 Authorization and ICNs – Beyond Open Content

Elwyn Brian Davies (Trinity College Dublin, IE)

ICNs to date are primarily focussed on open access infomation. This presentation is intended to promote a discussion on applying authorization in order to restrict access to particular content that is published into an ICN.

3.5 ICN Service Model Issues. Which are the ICN services, who provides them, and what does an ICN API look like?

Anders Eriksson (Ericsson Research – Stockholm, SE)

Joint work of Eriksson, Anders; Ohlman, Börje; Persson, Karl-Åke
Main reference A. Eriksson, B. Ohlman, K.-A. Persson, "What are the Services of an Information-centric Network, and Who Provides Them?", in: Proc. of the Fourth Int'l Conf. on Advances in P2P Systems (AP2PS 2012), IARIA, 2012.
URL http://www.thinkmind.org/index.php?view=article&articleid=ap2ps_2012_1_20_30008

This presentation highlights the need to define ICN services in addition to a basic content retrieval service. The following services are proposed: advertise, publish, subscribe, event notification, and search. The event notification service can alert the search service when new data is advertised or published, so that the search service can index the data in a timely fashion. Finally, an example of an ICN API is described.

3.6 ICN use cases and deployment (from device manufacturer perspective)

Myeong-Wuk Jang (Samsung, KR)

This talk includes ICN use cases and deployment issues from device manufacturer perspective. Although fundamental researches for ICN are very important, we should seriously consider CCN products that will be launched in the near future. The first product of ICN may not be networking devices, such as routers, but consumer devices, such as phones, computers, or TV. ICN should run over IP for a while. When we consider near future scenarios with ICN, there are still unresolved problems. These problems should be the target of our research and development.

3.7 Deploying ICN at the network edge

Vikas Kawadia (BBN Technologies – Cambridge MA, US)

Joint work of Kawadia, Vikas; Hoon, Jeremy; Parkes, David C.

We discuss the issues in deploying ICN at the mobile edge of the network. We consider incentives for users to participate in an ICN, in particular saving broadband quota by obtaining content locally from other users. We consider a simplified case where multiple users are competing for wireless bandwidth. We design a mechanism for incentive compatible dynamic prioritization of user data on shared routers, such as 3G/4G devices. We adopt a recent idea proposed by Babaioff et al. (2010) in a multi-armed bandits context to create a mechanism that is truthful for buyers, meaning that users or user devices can bid straightforwardly, and design a revenue pooling method for incentive alignment that makes the scheme faithful for sellers, meaning that sellers maximize revenue by using prioritization algorithms that preserve buyer truthfulness. The mechanism works for constant per-byte buyer values, private to each buyer, a wide range of stochastic demand models, and is ?detail-free,? in that the rules operate without knowledge of demand models or network conditions. We show simulation results demonstrating the efficiency gains from dynamic prioritization, as well as the effectiveness of revenue pooling.

3.8 Resource management in ICN: Business relations and interconnection

Luca Muscariello (Orange, FR)

In the talk we review the content-delivery value chain and how new business models, heavily driven by video delivery, are changing the interconnection relations between the different entities in the Internet. The talk is based on the viewpoint of a large network operator but also considers ISPs in general and their potential strategies in the long term with respect to the content-delivery value chain.

3.9 Video Streaming In NDN Architectures

Ashok Narayanan (Cisco Systems – Lexington, US)

This talk describes some of the issues seen with video streaming when implemented as an NDN application. It touches on the various service-providervideo streaming schemes (MPEG/UTP, progressive/RTMP and HTTP/adaptive), and describes some of the high-level issues when trying to re-implement HTTP adaptive streaming as an NDN application. It goes in-depth into bandwidth estimation which is a key feature of adaptive video streaming but is complicated by ICN/NDN architectures. One of the general lessons to be learned is

that applications which try to adapt to "network conditions" require significant changes for NDN/ICN because the representation of network conditions is quite different.

3.10 Blackadder: Node Design for an Information-Centric Network Architecture

George Parisis (University of Cambridge, GB)

Joint work of Parisis, George; Trossen, Dirk
Main reference D. Trossen, G. Parisis, "Designing and Realizing an Information-Centric Internet," in IEEE Communications Magazine, Volume 50, Issue 7, pp. 60–67, July 2012.
URL http://dx.doi.org/10.1109/MCOM.2012.6231280

Information-centric networking has been touted as an alternative to the current Internet architecture by several research groups. Our work addresses a crucial part of such a proposal, namely the design of a network node within an information-centric networking architecture. We describe the service model exposed to applications and other network nodes and we demonstrate a video streaming application that runs natively on top of a VPN network consisting of VMs across all European partners of the PURSUIT EU FP7 project.

3.11 Backscatter from the Data Plane – Threats to Stability and Security in Information-Centric Networking

Thomas C. Schmidt (HAW – Hamburg, DE)

Joint work of Schmidt, Thomas C.; Wählisch, Matthias; Vahlenkamp, Markus
Main reference M. Wählisch, T.C. Schmidt, M. Vahlenkamp, "Backscatter from the Data Plane – Threats to Stability and Security in Information-Centric Networking," arXiv:1205.4778v2 [cs.NI].
URL http://arxiv.org/abs/1205.4778v2

Information-centric networking proposals attract much attention in the ongoingsearch for a future communication paradigm of the Internet. Replacing the host-to-host connectivity by a data-oriented publish/subscribe service eases content distribution and authentication by concept, while eliminating threats from unwanted traffic at an end host as are common in today's Internet. However, current approaches to content routing heavily rely on data-driven protocol events and thereby introduce a strong coupling of the control to the data plane in the underlying routing infrastructure. In this paper, threats to the stability and security of the content distribution system are analyzed in theory and practical experiments. We derive relations between state resources and the performance of routers and demonstrate how this coupling can be misused in practice. We discuss new attack vectors present in its current state of development, as well as possibilities and limitations to mitigate them.

3.12 ICN evaluation: Why and How?

George Xylomenos (Athens University of Economics and Business, GR)

ICN evaluation seems to be very tricky: common networking metrics seem insufficient, existing traffic and network models are likely inappropriate and there qualitative arguments both for and against ICN. The real question though is what are we trying to prove? If existing metrics seem wrong, maybe this is because we expect ICN to offer something other than improved throughput or reduced delay. Whatever that is, it must be something to do with data naming.

3.13 What will be Inter-domain Policy in Content Centric Networks?

Eiko Yoneki (University of Cambridge, UK)

Inter-domain routing with BGP is not shortest paths, which involves policies on routing. In Information Centric Networking (ICN), it requires a policy based routing protocol with a finer granularity of policies (i.e. content names level rather than hosts). DiBenedetto et al. [1] explored routing policies in ICN. However, we have not seen much research in Inter-domain policy in ICN.

In ICN, the policies could be driven by economic incentives by deployment of sharing cache among peers, rebating routing, and multi-payment in multi-path, which could be complex. Rajahalme et al worked on incentive-compatible caching and peering in [2]. Agyapong et al also described implication for protocol design and public policy in economic incentives [3]. An economic model of ICN, where various stake holders play a complex game based on their incentives is an essential issue in ICN research. For example, content cache placement could depend on the inter-domain routing policy. This problem is important and interesting since no longer can the policy be operated on top of network protocols (i.e. TCP/IP), and it needs to be embedded within ICN.

References

1 S. DiBenedetto, C. Papadopoulos, and D. Massey, *Routing policies in named data networking*, in Proc. of the ACM SIGCOMM workshop on Information-centric networking (ICN), 2011

2 J. Rajahalme, M. Sarela, P. Nikander, and Sasu Tarkoma, *Incentive-compatible caching and peering in data-oriented networks*, ReArch 2008.

3 P. Agyapong and M. Sirbu, *Economic Incentives in Content-Centric Networking: Implications for Protocol Design and Public Policy*, 39th Research Conference on Communication, Information and Internet Policy, 2011.

3.14 Named Data Networking: Experimentation with the new architecture

Lixia Zhang (Univ. California – Los Angeles, US)

Joint work of the NDN project team
Main reference The Named Data Networking (NDN) Project, NDN Technical Report NDN-0001, October 2010.
URL http://named-data.net/

This talk reports on the progress of our NDN testbed development. Since its launch two years ago, the Named-Data Networking project has focused the research effort on the design, development, and actual usage of a variety of applications running over NDN networks. The NDN testbed is used not only to support application experimentations but also to drive research into network routing, monitoring, and management under the NDN architecture. Our experience suggests that architecture design research must include experimental components in order to verify whether, and how well, the proposed design can solve real problems. Our experience also shows that solving real problems not only forces architectural details to be filled in, but also validates and shapes the direction of the architectural development.

4 Working Groups

In this seminar, we organized two sessions for discussions in smaller groups. The first session covered two topics. Each topic has been analyzed by two individual groups. The second session focused on three different topics. In the following, we present the main results.

4.1 Applications & API (1)

Participants

Carsten Bormann, Antonio Carzaniga, Lars Eggert, Anders Eriksson, Xiaoming Fu, Volker Hilt, Pan Hui, Anders Lindgren, Ignacio Solis, George Xylomenos

Discussion and Open Questions

Starting from the key question what an API is, the group discussed where the API is located. In general, one may distinguish between a library API (running at the edge) and a core API (running everywhere). The group further discussed functions of a common API. This includes (a) advertise a name, (b) make data available, and (c) get data based on name. Requesting data might be split into `get` and `subscribe`. `Get` asks the network to retrieve a piece or a set of existing data based on a name. In contrast to this, `subscribe` asks the network to retrieve the data for a give amount of time.

Other topics that need consideration in the future are:

- Is there an Object Oriented API?
- Should CCN add an official way to advertise content?
- Are (CCN) conventions protocols?
- What about streams/flows?

4.2 Applications & API (2)

Participants

Marcus Brunner, Antonio Carzaniga, Elwyn Davies, Myeong-Wuk Jang, Gunnar Karlsson, Ashok Narayanan, Börje Ohlman, Lixia Zhang

Discussion and Open Questions

Any API discussion needs to clarify at which level the API will be deployed (application versus host stack API; host versus network API).

Raising the basic question whether we can define a common API across different ICN technologies, the group concludes that this seems not possible. Even basic verbs are not common across different approaches (e.g., *publish* is not relevant in NDN, *advertises* are not relevant in NetInf). Even an API for each individual ICN scheme is challenging as a number of subtleties are not fully defined. At the current state, we do not understand enough about our applications to able to define these APIs yet. The consensus of the group on this topic is that a *complete* API definition may be premature, but we need to keep working at our current model to harden it.

In the subsequent discussion, the group analyzed (a) *search* as a network primitives, (b) ICN as a network storage, (c) content revocation, and (d) interactive services. Only little consensus was found. Regarding the question: "Can ICN offer reliable, persistent network storage?", there are two observations. First, there is no guaranteed reliability without explicit agreement. Second, ICN offers ability to delegate storage. All that is missing is some protocol behaviour to bridge the gap between agreement and publication. This can be implemented by PubSub and with NDN.

Regarding content revocation they argue that this is not likely in any case. It is also not apparently a good use of anybody's time. Do something simple (like do-not-cache bits) and the rest of the problem belongs to the application.

In addition to this, the following open questions remain:

- Are networks sender driven or receiver driven?
- How do you implement a push model in NDN?
- API for "the network cannot find this"?
- Security?

4.3 Metrics and Evaluation (1)

Participants

Giovanna Carofiglio, Patrick Crowley, Van Jacobson, Vikas Kawadia, Dirk Kutscher, Luca Muscariello, George Parisis, Christian Esteve Rothenberg, Thomas C. Schmidt, Eiko Yoneki

Discussion and Open Questions

The main topics of this group includes (a) project testbeds, (b) metrics, and (c) economic aspects. The group noted the importance of controllable testbeds, i.e., dedicated hardware resources for repeatable experiments.

For an exhaustive list of metrics, the group pointed to the presentation given by Giovanna Carofiglio. They highlighted that the benefits of low-level caching, which helps in case of

packet loss (retransmission) and mobility scenarios. Memory can be used to implement three things: rate adaptation, repairing, and re-use (sharing).

An additional focus in the context of metric discussion was forwarding properties and congestion control. A major insight lies in the consideration of RTT variation. A common assumption is that spectrum of RTTs in CCN is lower. Other issues regard fairness with locally cached content (low RTT) and far-away content (high RTT). In particular, a CCN application cannot know where named pieces come from.

Regarding economic aspects, there was no agreement on the revenue model (i.e., which stakeholder gets which value). Metrics for network management and OPEX profiles should also be part of future work.

4.4 Metrics and Evaluation (2)

Participants

Bengt Ahlgren, Somaya Arianfar, Ken Calvert, Kevin Fall, Holger Karl, Pasi Sarolahti, Matthias Wählisch

Discussion and Open Questions

The group agreed that concrete testbeds should be preferred in contrast to generic testbeds. Then, the objectives of a testbed have been discussed. It is impossible to *verify* that an architecture is good. It is doable to *falsify* that an architecture is good. However, the most reasonable intention behind a testbed is to get concepts *running*.

A major open question is: How to quantify "betterness"?

The group came up with only small additions to Giovanna's list of metrics. This includes ease of programmability, accountability, ease of introducing net business models. The group noted that performance evaluation will only give arguments to dismiss ICN. Hence, we need elaborate arguments beside performance evaluation to push ICN.

Discussions have emphasized that limiting ICN to caching will miss 90% of today's applications.

4.5 ICN vs. DTN

Participants

Bengt Ahlgren, Somaya Arianfar, Patrick Crowley, Elwyn Brian Davies, Kevin Fall, Pan Hui, Gunnar Karlsson, Vikas Kawadia, Dirk Kutscher, Anders Lindgren, Pasi Sarolahti, Ignacio Solis, Eiko Yoneki

Discussion and Open Questions

After clarifying basic terms Delay Tolerant Networking (DTN) and ICN, the group presented a nice table comparing both approaches. The group concluded that ICN and DTN are not redundant but similar. Both schemes generally can provide disruption tolerant operations for link outages. There are two essential requirements for DTN support, location-independent naming and node with storage, which is shared with ICN.

4.6 Interdomain routing, forwarding

Participants

Ken Calvert, Giovanna Carofiglio, Antonio Carzaniga, Xiaoming Fu, Volker Hilt, Holger Karl, Luca Muscariello, Ashok Narayanan, Borje Ohlman, George Parisis, Christian E. Rothenberg, Thomas C. Schmidt, Matthias Wählisch, George Xylomenos

Discussion and Open Questions

The intention of this group meeting was to discuss both, interdomain routing and forwarding. However, The time was too short to discuss forwarding aspects.

The routing topic splits basically into two parts: (a) routing on topology-independent labels and (b) routing on topology-dependent labels.

For the first, there are two broad options. Distributing routing tables based on topology-independent (TI) names, and converting topology-independent names to topology-dependent (TD) addresses via lookup schemes. Routing tables and forwarding are then based on TD addresses. Obviously, hybrid concepts may also exist. Still an open question: Can inter-domain routing scale based on topology-independent labels? Even if we currently cannot imagine routing schemes that can solve it, it does not mean it cannot be solved.

Regarding routing on topology-dependent labels there was also no real consensus. Apart from mobility aspects people did not seem interested.

Overall, the participants concluded that a tutorial on network theory may help improve the debate about TI ot TD labels.

After this, the extension of interdomain routing policy has been discussed. It is quite likely that this will happen in the future as CDNs already exchange much more sophisticated policies. It has been identified as a useful area of future research.

Further open questions are

- Can incomplete-table type routing schemes work?
- Can proposed lookup service store # of states

4.7 ICN Deployment

Participants

Carsten Bormann, Marcus Brunner, Lars Eggert, Anders Eriksson, Myeong-Wuk Jang, Lixia Zhang

Discussion and Open Questions

Starting from the observation that ICN is easy to deploy as an overlay or in the infrastructure, it is still not fully clear why people would want to do this. In other words, we still miss a killer application.

As a disruptive technology arising that we can piggyback ICN on, the following has been noted. First, a big player that could make ICN actually happen is Google with Android. However, why would they want to do this. Elaborating this could be part of future work.

Another interesting application field for ICN may also machine-to-machine communication and the Internet of Things (IoT). If applying ICN to IoT, it will give us some sane routing (compared to the current protocols pushed in the IETF), which could be a killer app.

Participants

- Bengt Ahlgren
Swedish Institute of Computer Science – Kista, SE
- Somaya Arianfar
Aalto University, FI
- Carsten Bormann
Universität Bremen, DE
- Marcus Brunner
Swisscom AG – Bern, CH
- Ken Calvert
University of Kentucky, US
- Giovanna Carofiglio
ALCATEL – Marcoussis, FR
- Antonio Carzaniga
University of Lugano, CH
- Patrick Crowley
Washington University, US
- Elwyn Brian Davies
Trinity College Dublin, IE
- Lars Eggert
NetApp Deutschland GmbH – Kirchheim, DE
- Anders Eriksson
Ericsson Res. – Stockholm, SE
- Christian Esteve Rothenberg
CPqD – Campinas, BR
- Kevin Fall
Qualcomm Corp. – Berkeley, US
- Xiaoming Fu
Universität Göttingen, DE
- Volker Hilt
Alcatel-Lucent, DE
- Pan Hui
TU Berlin, DE
- Van Jacobson
PARC – Palo Alto, US
- Myeong-Wuk Jang
Samsung, KR
- Holger Karl
Universität Paderborn, DE
- Gunnar Karlsson
KTH – Stockholm, SE
- Vikas Kawadia
BBN Technologies – Cambridge MA, US
- Dirk Kutscher
NEC Laboratories Europe – Heidelberg, DE
- Anders Lindgren
Swedish Institute of Computer Science – Kista, SE
- Luca Muscariello
Orange Labs, FR
- Ashok Narayanan
Cisco Systems – Lexington, US
- Börje Ohlman
Ericsson Res. – Stockholm, SE
- George Parisis
University of Cambridge, GB
- Pasi Sarolahti
Aalto University, FI
- Thomas C. Schmidt
HAW – Hamburg, DE
- Ignacio Solis
PARC – Palo Alto, US
- Matthias Wählisch
FU Berlin, DE
- George Xylomenos
Athens University of Economics and Business, GR
- Eiko Yoneki
University of Cambridge, GB
- Lixia Zhang
Univ. California – Los Angeles, US

Report from Dagstuhl Seminar 12362

The Multilingual Semantic Web

Edited by

Paul Buitelaar[1], Key-Sun Choi[2], Philipp Cimiano[3], and Eduard H. Hovy[4]

1 **National University of Ireland – Galway, IE,** paul.buitelaar@deri.org
2 **KAIST – Daejeon, KR,** kschoi@kaist.ac.kr
3 **Universität Bielefeld, DE,** cimiano@cit-ec.uni-bielefeld.de
4 **University of Southern California – Marina del Rey, US**

Abstract

This document constitutes a brief report from the Dagstuhl Seminar on the "Multilingual Semantic Web" which took place at Schloss Dagstuhl between September 3rd and 7th, 2012 [1]. The document states the motivation for the workshop as well as the main thematic focus. It describes the organization and structure of the seminar and briefly reports on the main topics of discussion and the main outcomes of the workshop.

Seminar 02.–09. September, 2012 – www.dagstuhl.de/12362
1998 ACM Subject Classification H.1.2 User/Machine Systems: Human Factors, H.2.3 Languages, H.5.2 User Interfaces: Standardization
Keywords and phrases Semantic Web, Multilinguality, Natural Language Processing
Digital Object Identifier 10.4230/DagRep.2.9.15

1 Executive Summary

Paul Buitelaar
Key-Sun Choi
Philipp Cimiano
Eduard H. Hovy

The amount of Internet users speaking native languages other than English has seen a substantial growth in recent years. Statistics from 2010 in fact show that the number of non-English Internet users is almost three times the number of English-speaking users (1430 million vs. 536 million users). As a consequence, the Web is turning more and more into a truly multilingual platform in which speakers and organizations from different languages and cultural backgrounds collaborate, consuming and producing information at a scale without precedent. Originally conceived by Tim Berners-Lee et al. [5] as "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation", the Semantic Web has seen an impressive growth in recent years in terms of the amount of data published on the Web using the RDF and OWL data models. The kind of data published nowadays on the Semantic Web or Linked Open Data (LOD) cloud is mainly of a factual nature and thus represents a basic body of knowledge

[1] Please also visit the website http://www.dagstuhl.de/12362

The Multilingual Semantic Web, *Dagstuhl Reports*, Vol. 2, Issue 9, pp. 15–94
Editors: Paul Buitelaar, Key-Sun Choi, Philipp Cimiano, and Eduard H. Hovy
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

that is accessible to mankind as a basis for informed decision-making. The creation of a level playing field in which citizens from all countries have access to the same information and have comparable opportunities to contribute to that information is a crucial goal to achieve. Such a level playing field will also reduce information hegemonies and biases, increasing diversity of opinion. However, the semantic vocabularies used to publish factual data in the Semantic Web are mainly English, which creates a strong bias towards the English language and culture. As in the traditional Web, language represents an important barrier for information access as it is not straightforward to access information produced in a foreign language. A big challenge for the Semantic Web therefore is to develop architectures, frameworks and systems that can help in overcoming language and national barriers, facilitating the access to information originally produced for a different culture and language. An additional problem is that most of the information on the Web stems from a small set of countries where majority languages are spoken. This leads to a situation in which the public discourse is mainly driven and shaped by contributions from those countries where these majority languages are spoken. The Semantic Web vision bears an excellent potential to create a level playing field for users with different cultural backgrounds, native languages and originating from different geo-political environments. The reason is that the information available on the Semantic Web is expressed in a language-independent fashion and thus bears the potential to be accessible to speakers of different languages if the right mediation mechanisms are in place. However, so far the relation between multilingualism and the Semantic Web has not received enough attention in the research community. Exploring and advancing the state-of-the-art in information access to the Semantic Web across languages is the goal of the seminar proposed here. A Semantic Web in which information can be accessed across language and national barriers has important social, political and economic implications:

- it would enable access to data in other languages and thus provide support for direct comparisons (e.g. of public spending), thus creating an environment where citizens feel well-informed and contributing to increasing their trust and participation in democratic processes as well as strengthening democracy and trust in government and public administration
- it would facilitate the synchronization and comparison of information and views expressed in different languages, thus contributing to opinion forming processes free of any biases or mainstream effects
- it would foster higher information transparency; the exchange of many data items is limited due to national boundaries and national idiosyncrasies, as it is e.g. the case with financial data, the exchange of which is limited due to the availability of very different accounting procedures and reporting standards. Creating an ecosystem in which financial information can be integrated across countries can contribute to a higher transparency of financial information, global cash flow and investments.

Vision, Goals and Topic: The vision underlying the proposed workshop is the creation of a Semantic Web in which all languages have the same status, every user can perform searches in their own language, and information can be contrasted, compared and integrated across languages. As a main topic for the seminar, we intend to discuss in how far the Semantic Web can be extended —- from an infrastructural and conceptual point of view — in order to support access across languages. This will lead us to the discussion of two main questions:

- Ontological vocabularies that are available and used in the Semantic web cover a broad number of domains and topics to varying degrees of detail and granularity. For one

thing we will discuss in how far these vocabularies can indeed be seen as an interlingua (language-independent) representation. This includes the question how, building on such an interlingual representation, the Semantic Web can indeed support access to semantic data across languages. This discussion will extend to the question which approaches are suitable to translate the user's information needs, expressed in natural language, into such a language-independent representation.
- For another thing, we will discuss how the multilingual Semantic Web can be constructed by publication and linking of available multilingual lexical resources following the Linked Data paradigms. In this context, we will also discuss how natural language processing tools can benefit from such a linked ecosystem of lexico-semantic background knowledge.

Other topics that we anticipated would be discussed at the seminar include the following:

- models for the integration of linguistic information with ontologies, i.e., models for multilingualism in knowledge representation, in particular OWL and RDF(S)
- collaborative design of ontologies across languages and cultures
- multilingual ontology alignment
- multilingual and cross-lingual aspects of semantic search and querying of knowledge repositories
- cross-lingual question answering over Linked Data
- architectures and infrastructure for a truly Multilingual Semantic Web
- localization of ontologies to multiple languages
- automatic integration and adaptation of (multilingual) lexicons with ontologies
- multi- and cross-lingual ontology-based information extraction and ontology population
- multilingualism and linked data (generation, querying, browsing, visualization and presentation)
- multilingual aspects of ontology verbalization
- ontology learning across languages
- NLP methods to construct the multilingual Semantic Web

Organization & Structure

The Dagstuhl seminar on the Multilingual Semantic Web took place at Schloss Dagstuhl from the 3rd to the 7th of September 2012. The organizers were Paul Buitelaar (National University of Ireland, Galway), Key-Sun Choi (KAIST), Philipp Cimiano (Bielefeld University) and Eduard Hovy (CMU).

The organizers asked participants to submit an abstract and to prepare a short presentation of about 10 minutes for the seminar. The schedule of the seminar proposed by the organizers was as depicted in the figure below:

	Day 1	Day 2	Day 3	Day 4	Day 5
9:00 – 10:30	Introduction	Group Reports	Group Reports	Group Reports	Group Reports
10:30 – 12:00	Panel 1	Panel 2	Panel 3	Panel 4	Wrap-up Discussion
12:00 – 13:30	Lunch				
13:30 – 15:30	Group Work/Discussion				
15:30 – 16:00	Coffee Break				
16:00 – 17:00	Group Writing/Summarizing				
17:00 – 18:00	EU Perspective Kimmo Rossi	Industry & Application Perspective van Grondelle, Lieske & Sasaki	Walking	Demo Session *Contact Organizers*	
18:00	Dinner				

Panel 1	NLP4SW, SW4NLP	Sebastian Hellman, Graeme Hirst, Roberto Navigli, Sergei Nirenburg, Hans Uszkoreit
Panel 2	Multilingual Linked Data & Language Resources	Nicoletta Calzolari, Christian Chiarcos, Asun Gomez-Perez, Nancy Ide, John McCrae, Martin Volk, Ernesto Willem de Luca, Gerard de Melo
Panel 3	Multilinguality, Diversity & Collaboration	Dimitra Anastasiou, Bo Fu, Chu-Ren Huang, Antoine Isaac, Laurette Pretorius, Gabriele Sauberer, Gerard Budin
Panel 4	Multilingual Web Content Processing	Pushpak Bhattacharyya, Manuel Carrasco Benite, Iryna Gurevych, Aarne Ranta, Josef van Genabith

The first day started with an introduction by the organizers, giving an overview of the main topics and goals of the seminar. Some guiding questions for the seminar as proposed by the organizers were the following:

- Can we exploit the LOD for NLP?
- Can we allow for multilingual access to the knowledge in the LOD?
- Can we regard the LOD as an interlingua?
- Can we apply Linked Data principles to the modelling of linguistic/lexical resources?
- How can we facilitate the localization of (semantic) web sites to multiple languages?

As technical and research challenges for the field in the next years, the organizers highlighted the following:

- Aggregating and summarizing content across languages
- Repurposing and verbalizing content in multiple languages
- Linking of information across languages
- Detection of inconsistent views across languages
- Translation of "objects" that have a context and are produced within some workflow
- Large-scale and robust text analysis in multiple languages
- Personalized and contextualized Interpretation of NL [38]
- Cross-lingual/cultural reconciliation of conceptualizations

Every day, between 10:30 and 12:00, a panel took place in which attendees of the seminar had 10 minutes to present their view on the main challenges in the field, answering to the following questions in particular:

1. What are in your view the most important challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web?

2. Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?
3. Which figures are suited to quantify the magnitude or severity of the problem?
4. Why do current solutions fail short?
5. What insights do we need in order to reach a principled solution? What could a principled solution look like?
6. How can standardization (e.g. by the W3C) contribute?

After each panel the organizers attempted to group participants into teams around a certain topic. The groups worked together on the topic in the afternoons between 13:30 and 15:30. They were supposed to wrap-up their discussion and come up with a summary of their discussion until 17:00. These summaries were then presented in a plenary session to all the participants from Tuesday to Friday between 9:00 and 10:30.

Every day between 17:00 and 18:00 (just before dinner), we had an invited talk or special activity. On the first day, Kimmo Rossi from the European Commission shared his perspective on the challenges in our field. On the second day, there was a non-academic slot: First Jeroen van Grondelle showcased an industrial application of semantic, multilingual technologies; next, Christian Lieske and Felix Sasaki discussed perception and reality of the multilingual Semantic Web. On the third day we had a small walk to Noswendel (see Figure 1), and on the fourth day we organized a demo session, giving participants the opportunities to give a hands-on look at their tools.

Figure 1 Walking to Noswendel.

2 Table of Contents

3 Overview of Talks

3.1 Some reflections on the IT challenges for a Multilingual Semantic web

Guadalupe Aguado de Cea and Elena Montiel Ponsoda (Universidad Politécnica de Madrid)

1. **Most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web (SW):**
   Many attempts have been made to provide multilinguality to the Semantic Web, by means of annotation properties in Natural Language (NL), such as RDFs or SKOS labels, and other lexicon-ontology models, such as lemon, but there are still many issues to be solved if we want to have a truly accessible Multilingual Semantic Web (MSW). Reusability of monolingual resources (ontologies, lexicons, etc.), accessibility of multilingual resources hindered by many formats, reliability of ontological sources, disambiguation problems and multilingual presentation to the end user of all this information in NL can be mentioned as some of the most relevant problems. Unless this NL presentation is achieved, MSW will be restricted to the limits of IT experts, but even so, with great dissatisfaction and disenchantment.
2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?**
   Considering Linked Data as a step forward from the original Semantic Web, providing the possibility of accessing all the information gathered in all the ontological resources should become one significant objective, if we want every user to "perform searches in their own language", as mentioned in the motivation of Dagstuhl Seminar. Globalization of work has opened the scope of possible domains and sectors interested in Linked data and a true MSW. From governmental, political, administrative and economic issues to medicine, chemistry, pharmaceutical, car makers and other industries alike, all would hop on the bandwagon of MSW if it provides them the suitable information needed for their businesses. As long as we cannot retrieve the answer to a question in NL, even if we have the possible information in DBpedia and other ontological and knowledge resources, it will be difficult to beat Google, and extract the most of LD and the SW, no matter how many "semantic" resources we have.
3. **Which figures are suited to quantify the magnitude or severity of the problem?**
   It is difficult for us to quantify the problem in figures, but it is clear that we can miss the boat if this issue remains unsolved. In the last few years the mobile industry has made advances at a greater speed, maybe because there were more chances to make money.
4. **Why do current solutions fail short?**
   At the moment, we have complex models to be implemented by SW illiterate, many technological issues unsolved, and a lack of agreement with respect to the ontological-lexical linguistic knowledge to be provided to end-users when using the SW to improve their resources.
5. **What insights do we need in order to reach a principled solution? What could a principled solution look like?**
   Focusing on certain aspects that can be agreed upon by many key sectors (researchers, developers, industry, end-users), some relevant problems could be approached aiming

at delimiting the wishes, needs and resources available. A principled solution should be based on simplicity, usefulness, wide coverage, and reusability.

6. **How can standardization (e.g. by the W3C) contribute?**
   It can contribute because participation is open to many sectors involved. If all sectors cooperate, dissemination and promotion can be achieved more easily. Getting other standardization committees involved (ISO TC 37) can also widen the scope and can contribute to dissemination too. But it is important to get industry professionals involved to make them aware of the possibilities they have to make the most of their products.

3.2 Accessibility to a Pervasive Web for the challenged people

Dimitra Anastasiou (University of Bremen)

1. **What are in your view the most important challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web?**
   One need is to make people believe about its importance. Although some projects and workshops (including the Dagstuhl Workshop) bring this topic forward, there is still need for more interesting projects and initiatives in the community. As Semantic Web technologies are used by many domains, and multilingualism is also an aspect taken into account by many stakeholders, many people regard the Multilingual Semantic Web (MSW) as a vague concept, so some clear description, specifications or even a standard would make the MSW more prominent. At the moment I am more interested in the accessibility to the Web and the MSW by the seniors and people with disabilities. Moreover, and in relation to the Web for challenged people, I am interested in the pervasive Web in Ambient Assisted Living (AAL), which goes beyond the Web present on a PC monitor, and is present in the invisible technology in smart homes.
2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?**
   The aging phenomenon is reality today, as according to the World Population Aging report, the world average of the 65+ age group was 7.6% in 2010 and will be 8.2% in 2015. The European Commission suggests demographic and epidemiological research on aging and disability, predicting the size of the future aging population, and acquiring information as inputs to planning. Industries (and some academic groups) are mostly concerned with AAL, but the community researching on the Web technology used particularly there is very small. Moreover, multilingualism plays a secondary role, though it is so important, as seniors today are often not foreign language speakers and have to communicate with technology (Web or not). Whereas health informatics, HCI, HRI, sensoring and recognition play an important role, the Semantic Web and multilingual support are not taken into serious consideration.
3. **Which figures are suited to quantify the magnitude or severity of the problem?**
   The Working Draft of "Web Accessibility for Older Users: A Literature Review"[2] gives

[2] Web Accessibility for Older Users: A Literature Review: http://www.w3.org/TR/wai-age-literature/

very interesting insights about Web design and development and its aspects affecting the elderly.

4. **Why do current solutions fail short?**
   Because the limitations of those challenged people can vary significantly, it cannot be really categorized in specific groups, so high customization of software and high learning effort is needed, which results in information overload. The technology is too expensive and not affordable yet. Moreover, it is also very complex, so easier-to-use and user-friendly methods should be developed.
5. **What insights do we need in order to reach a principled solution? What could a principled solution look like?**
   More initiatives including common projects, community groups workshops in the fields of AAL, multimodality, Semantic Web, language technology. A principled solution should look like elderly persons being able to speak in their mother tongue to turn on and off their coffee machine, switch on and off lights. When they speak in their mother tongue, they do not feel digitally intimidated, but are more natural, trustful, and user-friendly. Ontologies could help dialogue systems triggering predictable actions in AAL smart homes, i.e. turning off the oven when not used or reminding a person to make a phone call.
6. **How can standardization (e.g. by the W3C) contribute?**
   Cooperation with the W3C Web Accessibility Initiative[3] would be very useful. It has released Web Content Accessibility Guidelines[4], User Agent Accessibility Guidelines, and Authoring Tool Accessibility Guidelines.

3.3 Multilingual Computation with Resource and Process Reuse

Pushpak Bhattacharyya (Indian Institute of Technology Bombay)

1. **Introduction**
   Mutilingual computation is the order of the day and is needed critically for the realization of the Semantic web dream. Now, it stands to reason, that work done for a language should come to help for computation in another language. For example, if through the investment of resources we have been able to detect named entities in one language, we should be able to detect them in another language too, through much smaller levels of investment like transliteration. The idea of projection from one language to another is a powerful and potent one and merits deep investigation. In the seminar I would like to expound on the projection for multilingual NLP.
2. **Challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web**
   Resource constraint is the most important challenge facing multilingual access to the Semantic web. Over the years through conscious decisions, English has built foundational tools and resources for language processing. Examples of these are Penn Treebank [5],

[3] W3C Web Accessibility Initiative (WAI): http://www.w3.org/WAI/
[4] Web Content Accessibility Guidelines (WCAG) Overview: http://www.w3.org/WAI/intro/wcag.php
[5] http://www.cis.upenn.edu/~treebank/

Propbank, Rule based and Statistical Parsers[6], Wordnet[7], Corpora of various kinds of annotation and so on and so forth. No language comes anywhere close to English in terms of lexical resources and tools.

3. **Why does the problem matter in practice?**
   It is impossible to do NLP without adequate lexical resources and foundational tools. For example, nobody thinks of building a parser today for a language, without first creating Treebank for the language – constituency or dependency – and then training a probabilistic parser on the Treebank. However, creating treebanks requires years of effort. Everything in language technology sector needs lexical resources. Information Extraction, Machine Translation and Cross Lingual Search are some of the examples. E-Governance – a domain dealing with the automatization of administrative processes of the government in a large, multilingual country like India – is a large consumer of language technology.
4. **Which figures are suited to quantify the magnitude or severity of the problem?**
   Lexical resources are typically quantified by the amount of annotated data and foundational tools by their precision and recall figures. For example, the famous SemCor [8] corpus for sense annotated data has about 100,000 Wordnet id marked words. On the tools side, CLAWS POS tagger for English has over 97% accuracy.
5. **Why do current solutions fail short?**
   It takes years to build high quality lexical resources. Both linguistic expertise and computational dexterity are called for. It is not easy to find people with both linguistic and computational acumen. Large monetary investment to is called for.
6. **Principled Solution**
   Projection is the way to go. Reuse of resources and processes is a must. Over the years in our work on word sense disambiguation involving Indian languages, we have studied how sense distributions can be projected from one language to another for effective WSD [47, 48, 50, 49]. The idea of projection has been applied in POS tagging (best paper award ACL 2011[9]). We have also used it to learn named entities in one language from the NE tagged corpora of another language.
7. **How can standardization (e.g. by the W3C) contribute?**
   For projection to work at all, resources and tools need to be standardized for input-output, storage, API and so on. For example, WordNet building activity across the world follows the standard set by the Princeton WordNet.

---

[6] http://nlp.stanford.edu/software/lex-parser.shtml
[7] http://wordnet.princeton.edu
[8] http://www.gabormelli.com/RKB/SemCor_Corpus
[9] Dipanian Das and Slav Petrov, Unsupervised Part-of-Speech Tagging with Bilingual Graph-based Projections (ACL11); Singapore, August, 2009

3.4 Multilingual Semantic Web and the challenges of Open Language Data

Nicoletta Calzolari (Istituto Linguistica Computazionale, Pisa)

Language Technology (LT) is a data-intensive field and major breakthroughs have stemmed from a better use of more and more Language Resources (LRs). LRs and Open/Shared Language Data is therefore a great topic! New approaches are needed, both for Data and Meta-Data (LRs and Meta- LRs). My topics are linked to the layer of LRs and language services that serve LT, and especially open information on LRs and on research results. How can Linked Data contribute?

1. **The Language Resource dimensions**
   In the FLaReNet[10] Final Blueprint, the actions recommended for a strategy for the future of the LR field are organised around nine dimensions: a) Infrastructure, b) Documentation, c) Development, d) Interoperability, e) Coverage, Quality and Adequacy, f) Availability, Sharing and Distribution, g) Sustainability, h) Recognition, i) International Cooperation. Taken together, as a coherent system, these directions contribute to a sustainable LR ecosystem. Multilingual Semantic Web has strong relations with many of these dimensions, esp. a), b), d), f), g).

2. **Language Resources and the Collaborative framework**
   The traditional LR production process is too costly. A new paradigm is pushing towards open, distributed language infrastructures based on sharing LRs, services and tools. It is urgent to create a framework enabling effective cooperation of many groups on common tasks, adopting the paradigm of accumulation of knowledge so successful in more mature disciplines, such as biology, astronomy, physics. This requires the design of a new generation of multilingual LRs, based on open content interoperability standards [12]. Multilingual Semantic Web may help in determining the shape of the LRs of the future, consistent with the vision of an open distributed space of sharable knowledge available on the web for processing (see [11]). It may be crucial to the success of such an infrastructure, critically based on interoperability, aimed at improving sharing of LRs and accessibility to multilingual content. This will serve better the needs of language applications, enabling building on each other achievements, integrating results, and having them accessible to various systems, thus coping with the need of more and more 'knowledge intensive' large-size LRs for effective multilingual content processing. This is the only way to make a great leap forward.

3. **Open Documentation on LRs**
   Accurate and reliable documentation of LRs is an undisputable need: documentation is the gateway to discovery of LRs, a necessary step towards promoting the data economy. LRs that are not documented virtually do not exist: initiatives able to collect and harmonise metadata about resources represent a valuable opportunity for the NLP community.

   LRE Map: The LRE Map is a collaborative bottom-up means of collecting metadata on LRs from authors. It is an instrument for enhancing availability of information about

[10] http://www.flarenet.eu

LRs, either new or already existing ones, and a way to show the current LR landscape and its trends. As a measuring tool for monitoring various dimensions of LRs across places and times, it helps highlighting evolutionary trends in LR use and related development by cataloguing not only LRs in a narrow sense (i.e. language data), but also tools, standards, and annotation guidelines. The Map contributes to the promotion of a movement towards an accurate and massive documentation of LRs.

4. **Open Language Resource Repositories**
The rationale behind the need of Open LR Repositories is that accumulation of massive amounts of (high-quality) multi-dimensional data about many languages is the key to foster advancement in our knowledge about language and its mechanisms. We must be coherent and take concrete actions leading to the coordinated gathering – in a shared effort – of as many (processed/annotated) language data as we are able to produce. This initiative compares to the astronomers/ astrophysics' accumulation of huge amounts of observation data for a better understanding of the universe.

   Language Library: The Language Library is an experiment – started around parallel/comparable texts processed by authors at LREC 2012 – of a facility for gathering and making available the linguistic knowledge the field is able to produce, putting in place new ways of collaboration within the community. It is collaboratively built by the community providing/enriching LRs by annotating/processing language data and freely using them. The multi-layer and multi-language annotation on the same parallel/comparable texts should foster comparability and equality among languages. The Language Library is conceived as a theory-neutral space, which allows for several annotation philosophies to coexist, but we must exploit the sharing trend for initiating a movement towards creating synergies and harmonisation among annotation efforts that are now dispersed. In a mature stage the Library could focus on enhancing interoperability, encouraging the use of common standards and schemes of annotation. Interoperability should not be seen as a superimposition of standards but rather as the promotion of a series of best practices that might help other contributors to better access and easily reuse the annotation layers provided. The Language Library could be seen as the beginning of a big Genome project for languages, where the community collectively deposits/creates increasingly rich and multi-layered LRs, enabling a deeper understanding of the complex relations between different annotation layers/language phenomena.

5. **Open Repositories of Research Results**
Disclosing data/tools related to published papers is another "simpler" addition to the Language Library, contributing to the promotion of open repositories of LR research results. Moreover LRs must be not only searchable/shareable, but also "citable" (linked to issue h) Recognition).

6. **Open Language Data (OpenLanD)**
Open Language Data – the set of 2. to 5. above – aims at offering the community a series of facilities for easy and broad access to information about LRs in an authoritative and trustable way. By investing in data reusability, OpenLanD can store the information as a collection of coherent datasets compliant to the Linked Data philosophy. The idea is that by linking these data among themselves and by projecting them onto the wider background of Linked Data, new and undiscovered relations can emerge. OpenLanD must be endowed with functionalities for data analytics and smart visualisation. OpenLanD

differs from existing catalogues for the breadth and reliability of information due to a community-based approach. The information made available covers usages, applications of LRs, their availability, as well as related papers, individuals, organisations involved in creation or use, standards and best practices followed or implemented. OpenLanD avoids the problem of rapid obsolescence of other catalogues by adopting a bottom-up approach to meta-data population.

3.5 Multilingual Web Sites

Manuel Tomas Carrasco Benitez (European Commission)

1. **Abstract**
   Multilingual Web Sites (MWS) refer to web sites that contain multilingual parallel texts; i.e., texts that are translations of each other. For example, most of the European Institutions sites are MWS, such as Europa [11]. The main point of views are:
   - Users should expect the same multilingual behaviour when using different browsers and/or visiting different web sites.
   - Webmasters should be capable of creating quickly high quality, low cost MWS.

   This is a position paper for the Dagstuhl Seminar on the Multilingual Semantic Web. Personal notes on this event can be found on the web[12].
2. **Relevance**
   MWS are of great practical relevance as there are very important portals with many hits; also they are very complex and costly to create and maintain: Europa is in 23 languages and contains over 8 million pages. Current multilingual web sites are applications incompatible with each other, so facilitating and enjoying this common experience entails standardisation. There is a Multilingual Web Sites Community Group at the W3C [13].
3. **Point of Views**
   a. **User** From a users point of view, the most common usage is monolingual, though a site might be multilingual; i.e., users are usually be interested in just one language of the several available at the server. The language selection is just a barrier to get the appropriate linguistic version. One has also to consider that some users might be really interested in several linguistic versions. It is vital to agree on common behaviours for users: browser-side (language button) and server-side (language page).
   b. **Webmaster** Webmaster refers to all the aspect of the construction of MWS: author, translator, etc. The objective is the creation of high quality low cost MWS. Many existing applications have some multilingual facilities and (stating the obvious) one should harvest the best techniques around. Servers should expect the same application programming interface (API). The first API could be just a multilingual data structure.

[11] Europa; http://europa.eu
[12] http://dragoman.org/dagstuhl
[13] Multilingual Web Sites Community Group; http://www.w3.org/community/mws

The absence of this data structure means that each application has to craft this facility; having the same data structure means that servers (or other programs) would know how to process this data structure directly. It is a case of production of multilingual parallel texts: the cycle Authorship, Translation and Publication chain (ATP-chain)[14].

4. **Wider context**
   - *Language vs. non-language aspects:* differentiate between aspects that are language and non-language specific. For example, the API between CMS and web server is non-language specific and it should be addressed in a different forum.
   - *Language as a dimension:* as in TCN [15], one should consider language a dimension and extend the con- cept to other areas such as linked data. Consider also feature negotiations as in TCN.
   - *Linguistic versions:* the speed (available now or later) and translation technique (human or machine translation) should be considered in the same model.
   - *Unification:* multilingual web is an exercise in unifying different traditions looking at the same object from different angles and different requirements. For example, the requirements for processing a few web pages are quite different from processing a multilingual corpus of several terabytes of data.
5. **Multidiscipline map**
   - Web technology proper
     - Content management systems (CMS), related to authoring and publishing
     - Multilingual web site (MWS)
     - Linked data, a form of multilingual corpora and translation memories
     - Previous versions in time, a form of archiving [16]
   - Traditional natural language processing (NLP)
     - Multilingual corpora, a form of linked data [17]
       - Machine translation, for end users and prepossessing translators
     - Source documents and tabular transformations, the same data in different presentations
   - Translation
     - Computer-aided translation (CAT)
       - Preprocessing, from corpora, translation memories or machine translation
     - Computer-aided authoring, as a help to have better source text for translation
     - Localisation
     - Translation memories (TM) [18], related to corpora and linked data
   - Industrial production of multilingual parallel publications
     - Integration of the Authorship, Translation and Publishing chain (ATP-chain)
     - Generation of multilingual publications
     - Official Journal of the European Union [19]
6. **Disclaimer**
   This document represents only the views of the author and it does not necessarily represent the opinion of the European Commission.

[14] Open architecture for multilingual parallel texts; http://arxiv.org/pdf/0808.3889
[15] Transparent Content Negotiation in HTTP; http://tools.ietf.org/rfc/rfc2295.txt
[16] Memento – Adding Time to the Web; http://mementoweb.org
[17] Multilingual Dataset Format; http://dragoman.org/muset
[18] TMX 1.4b Specification; http://www.gala-global.org/oscarStandards/tmx/tmx14b.html
[19] Official Journal of the European Union; http://publications.europa.eu/official/index_en.htm

3.6 The Multilingual Semantic Web and the intersection of NLP and Semantic Web

Christian Chiarcos (Information Sciences Institute, University of Southern California)

The premise of the Dagstuhl seminar is the question which problems we need to overcome in order to enhance multilingual access to the Semantic Web, and how these are to be addressed.

Ultimately, the Semantic Web in its present stage suffers from a predominance of resources originating in the Western hemisphere, with English as their primary language. Eventually, this could be overcome by providing translated and localized versions of resources in other languages, and thereby creating a critical mass of foreign language resources that is sufficient to convince potential non-English speaking users to (a) employ these resources, and (b) to develop their own extensions or novel resources that are linked to these. On a large scale, this can be done automatically only, comparable to, say, the conversion of the English Wikipedia into Thai[20]. Unlike the translation of plain text, however, this translation requires awareness to the conceptual structure of a resource, and is thus not directly comparable to text-oriented Machine Translation. A related problem is that the post-editing of translation results in a massive crowdsourcing approach (as conducted for the Thai Wikipedia) may be problematic, because most laymen will not have the required level of technical understanding.

Therefore, the task of resource translation (and localization) of Semantic Web resources requires a higher level of automated processing than comparable amounts of plain text. This is an active research topic, but pursued by a relatively small community. One possible issue here is that the NLP and Semantic Web communities are relatively isolated from each other[21], so that synergies between them are limited. A consequence is that many potentially interested NLP people are relatively unaware of developments in the Semantic Web community, and, moreover, that they do not consider Semantic Web formalisms to be relevant to their research. This is not only a problem for the progress of the Multilingual Semantic Web, but also for other potential fields of overlap. In the appendix I sketch two of them.

In my view, both the NLP community and the Semantic Web community could benefit from small- to mid-scale events co-located with conferences of the other community (or joint seminars, as this workshop), and that this may help to identify fields of mutual interest, including, among other topics, the translation of Semantic Web resources. In at least two other fields, such convergence processes may already be underway, as sketched below.

Questionnaire

1. **Challenges/ problems and needs with respect to the multilingual access to the Semantic Web**

   For languages that are under-represented in the Semantic Web, the initial bias to create resources in their own language and in accordance with their own culture is substantially higher than for English, where synergy effects with existing resources can be exploited in the development of novel resources. To provide these languages with a basic repository of

[20] http://www.asiaonline.net/portal.aspx#ThaiLaunch

[21] For example, the LREC (http://www.lrec-conf.org) lists 11 publications for the topic „Semantic Web“ for 2012,11 for 2010, 16 for 2008. Similarly, the Google counts for ACL (http://aclweb.org/anthology) contributions containing the word „ontology“ are consistently low: 2008: 5, 2009: 8, 2010: 15, 2011: 3, 2012: 7. Both conferences have between 500 and 1000 participants, so, in terms of paper-participant ratio, this line of research is underrepresented.

SW resources, massive automated translation is required. This task is, however, closer to the traditional realm of NLP than to that of the SW. The SW-subcommunity working towards this direction is thus relatively small, and may benefit from closer ties to the NLP community. (Which may be of mutual interest to both sides, also beyond the problem of Semantic Web multilingualism, see appendix.)

2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?**
   The situation is comparable to the development of NLP tools for less-resourced languages. Without a basic set of language- and culture-specific resources (say, a WordNet and a DBpedia/Wikipedia with sufficient coverage), there will be little interest to develop and to invest in Semantic Web applications. A plain translation is an important first step, but for semantic resources, there may be important culture-specific differences that need to be taken into consideration. These efforts can be crowd-sourced to a certain extent, but only if a certain level of knowledge is already available in order to convince contributors that this is an effort that pays off.
3. **Which figures are suited to quantify the magnitude or severity of the problem?**
   As for the primary problem to attract potentially interested NLP people, this can be illustrated by the small number of Semantic Web contributions to NLP conferences (and vice versa), see footnote 21.
4. **Why do current solutions fail short?**
   The NLP community and the SW community are relatively isolated from each other, and often not aware of developments in the other community. For example, a recent discussion on an NLP mailing list showed that occasionally RDF (as an abstract data model) is confused with RDF/XML (as one RDF linearization) and rejected because of the verbosity of this linearization, even though other, more compact and more readable linearizations exist.
5. **What insights do we need in order to reach a principled solution? What could a principled solution look like?**
   Co-located and/or interdisciplinary events. (Simply continue and extend the series of Multilingual Semantic Web and OntoLex workshops.) Interdisciplinary community groups.
6. **How can standardization (e.g. by the W3C) contribute?**
   Standardization is actually a key issue here. The NLP community developed its own standards within the ISO, and succeeded in integrating different groups from NLP/computational linguistics/computational lexicography. Semantic Web standards, however, are standardized by the W3C. Even though, say, GrAF and RDF (see appendix) are conceptually very close, the potential synergies have been realized only recently. If these standardization initiatives could be brought in closer contact with each other, natural convergence effects are to be expected.

Appendix

Possible Future Convergences between Semantic Web and NLP

From the perspective of Natural Language Processing and Computational Linguistics, one of the developments I would expect for the next 5-10 years is the accelerating convergence of both disciplines, at least in certain aspects. On the one hand, this includes adopting Linked Data as a representation formalism for linguistic resources in; on the other hand, this includes the improved integration of NLP tools and pipelines in Semantic Web applications. Both developments can be expected to continue for the next decade.

The Prospective Role of Linked Data in Linguistics and NLP

In the last 20 years, Natural Language Processing (NLP) has seen a remarkable maturation, evident, for example, from the shift of focus of shared tasks from elementary linguistic analyses over semantic analyses to higher levels of linguistic description [22]. To a large extent, this development was driven by the increased adaption of statistical approaches during the 1990s. One necessary precondition for this development was the availability of large-scale corpora, annotated for the phenomena under discussion, and for the development of NLP tools for higher levels of description (say, semantics or anaphoric annotation), the number and diversity of annotations available (and necessary) increased continually.

During the same period, corpus linguistics has developed into a major line of research in linguistics, partially supported by the so-called "pragmatic shift" in theoretical linguistics, when scholars have recognized the relevance of contextual factors. The study of these context factors favored the application of corpora in linguistics at a broader scale, which can now be considered to be an established research paradigm in linguistics.

Taken together, both communities created increasingly diverse and increasingly large amounts of data whose processing and integration, however, posed an *interoperability challenge*. In a response to this, the NLP community developed generic formalisms to represent linguistic annotations, lexicons and terminology, namely in the context of the ISO TC37. As far as corpora are concerned, a standard, GrAF [44], has been published this year. So far, GrAF is poorly supported with infrastructure and maintained by a relatively small community. However, its future application can take benefit of developments in the Linked Data community, where RDF provides a data model that is similar in philosophy and genericity, but that comes with a rich technological ecosystem, including data base implementations and query languages – which are currently not available for GrAF. Representing corpora in RDF, e.g., using an RDF representation of GrAF yields a number of additional benefits, including the uncomplicated integration of corpus data with other RDF resources, including lexical-semantic resources (e.g., WordNet) and terminology resources (e.g., GOLD). A comparable level of integration of NLP resources within a uniform formalism has not been achieved before, and to an increasing extent, this potential is recognized by researchers in NLP and linguistics, as manifested, for example, in the recent development of a Linguistic Linked Open Data cloud [23].

3.7 The importance of Semantic User Profiles and Multilingual Linked Data

Ernesto William De Luca (University of Applied Sciences Potsdam)

1. **Introduction**

Today, people start to use more and more different web applications. They manage their bookmarks in social bookmarking systems, communicate with friends on Facebook [24]

[22] CoNLL Shared Tasks: 1999-2003 flat annotations (NP bracketing, chunking, clause identification, named entity recognition), 2004-2009: dependency parsing and semantic role labelling, 2010-2012: pragmatics and discourse (hedge detection, coreference).
[23] http://linguistics.okfn.org/llod
[24] http://www.facebook.com/

and use services like Twitter [25] to express personal opinions and interests. Thereby, they generate and distribute personal and social information like interests, preferences and goals [68]. This distributed and heterogeneous corpus of user information, stored in the user model (UM) of each application, is a valuable source of knowledge for adaptive systems like information filtering services. These systems can utilize such knowledge for personalizing search results, recommend products or adapting the user interface to user preferences. Adaptive systems are highly needed, because the amount of information available on the Web is increasing constantly, requiring more and more effort to be adequately managed by the users. Therefore, these systems need more and more information about users interests, preferences, needs and goals and as precise as possible. However, this personal and social information stored in the distributed UMs usually exists in different languages (language heterogeneity) due to the fact that we communicate with friends all over the world.
Therefore, we believe that the integration of multilingual resources into a user model aggregation process to enable the aggregation of information in different languages leads to better user models and thus to better adaptive systems.

a. **The Use of Multilingual Linked Data**
Because the Web is evolving from a global information space of linked documents to one where both documents and data are linked, we agree that a set of best practices for publishing and connecting structured data on the Web known as Linked Data. The Linked Open Data (LOD) project [6] is bootstrapping the Web of Data by converting into RDF and publishing existing available "open datasets". In addition, LOD datasets often contain natural language texts, which are important to link and explore data not only in a broad LOD cloud vision, but also in localized applications within large organizations that make use of linked data [3, 66].
The combination of natural language processing and semantic web techniques has become important, in order to exploit lexical resources directly represented as linked data. One of the major examples is the WordNet RDF dataset [73], which provides concepts (called synsets), each representing the sense of a set of synonymous words [32]. It has a low level of concept linking, because synsets are linked mostly by means of taxonomic relations, while LOD data are mostly linked by means of domain relations, such as parts of things, ways of participating in events or socially interacting, topics of documents, temporal and spatial references, etc. [66].
An example of interlinking lexical resources like EuroWordNet[77] or FrameNet [26] [2] to the LOD Cloud is given in [19, 33]. Both create a LOD dataset that provides new possibilities to the lexical grounding of semantic knowledge, and boosts the "lexical linked data" section of LOD, by linking e.g. EuroWordNet and FrameNet to other LOD datasets such as WordNet RDF[73]. This kind of resources open up new possibilities to overcome the problem of language heterogeneity in different user models and thus allows a better user model aggregation [20].

2. **Requirements for a User-Oriented Multilingual Semantic Web**
Based on the idea presented above, some requirements have to be fulfilled:

[25] http://twitter.com/
[26] http://framenet.icsi.berkeley.edu/

Requirement 1: Ontology-based profile aggregation. We need an approach to aggregate information that is both application independent and application overarching. This requires a solution that allows to semantically define relations and coherences between different attributes of different UMs. The linked attributes must be easily accessible by applications such as recommender and information retrieval systems. In addition, similarity must be expressed in these defined relations.
Requirement 2: Integrating semantic knowledge. A solution to handle the multilingual information for enriching user profiles is needed. Hence, methods that incorporate information from semantic data sources such as EuroWordNet and that aggregate complete profile information have to be developed.

a. **Multilingual Ontology-based Aggregation**
For the aggregation of user models, the information in the different user models has to be linked to the multilingual information (as Multilingual Linked Data) as we want to leverage this information and use it for a more precise and qualitatively better user modeling. These resources can be treated as a huge semantic profile that can be used to aggregate user models based on multilingual information.
Figure 1 describes the general idea. The goal is to create one big semantic user profile, containing all information from the three profiles of the user information, were the data is connected. The first step is to add the multilingual information to the data contained in the different user models. This gives us a first model were the same data is linked together through the multilingual information.

b. **Integrating Semantic Knowledge**
The second step is to add links between data that is not linked through the multilingual information. The target is to have a semantic user model were data is not only connected on a language level, but also on a more semantic similarity level. The aggregation of information into semantic user models can be performed similarly to the approach described in [4], by using components that mediate between the different models and using recommendation frameworks that support semantic link prediction like [69]. The combined user model should be stored in an commonly accepted ontology, like [37], to be able to share the information with different applications.
With such a semantic user model, overcoming language barriers, adaptive systems have more information about the user and can use this data to adapt better to the user preferences.

3. **Conclusions**
Analyzing the problems described above, we believe that more context information about users is needed, enabling a context sensitive weighting of the information used for the profile enrichment. The increasing popularity of Social Semantic Web approaches and standards like FOAF [27] can be one important step in this direction. On the other hand, multilingual semantic datasets itself (as for example multilingual linked data) have to be enriched with more meta-information about the data. General quality and significance information, like prominence nodes and weighted relations, can improve semantic algorithms to better compute the importance of paths between nodes. Enriching the quality of user profiles and the multilingual semantic representation of data can

[27] http://www.foaf-project.org/

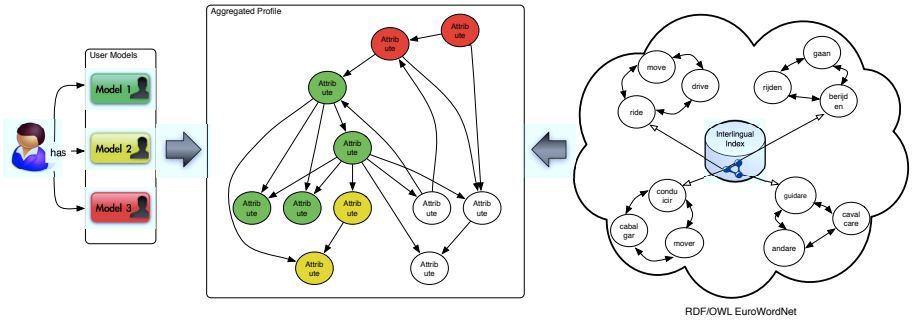


Figure 2 Integrating semantic knowledge about multilingual dependencies with the information stored in the user models.

be helpful, because both sides cover different needs required for an enhancement and consolidation of a multilingual semantic web.

3.8 Shared Identifiers and Links for the Linguistic Linked Data Cloud

Gerard de Melo (ICSI Berkeley)

The Web of Data opens up new opportunities in many areas of science and technology, including linguistics and library science. Common data formats and protocols have made it easier than ever to work with information from different sources simultaneously. The true potential of Linked Data, however, can only be appreciated when shared identifiers and extensive cross-linkage engender extensive interconnectedness across different data sets.

Examples of shared identifiers include those based on WordNet and Wikipedia. The UWN/MENTA multilingual knowledge base, for instance, integrates millions of words and names from different languages into WordNet and also uses Wikipedia-based identifiers [**?**]. This means that one buys into ecosystems already carrying a range of valuable pre-existing assets. WordNet, for instance, already comes with sense-annotated corpora and mappings to other resources. Wikipedia-based identifiers are also used by DBpedia [1], YAGO [41], and numerous other Linked Data providers.

Lexvo.org's linguistic identifiers are another example. Consider the example of a book written in a little-known under-resourced language. If its bibliographic entry relies on identifiers from Lexvo.org, one can easily look up where that language is spoken and what other libraries carry significant numbers of other books in the same language. Additionally, Lexvo.org also serves as an example of cross-linkage between resources. The service provides a language hierarchy [21] that connects identifiers based on the ISO 639 language standards to relevant entries in DBpedia, WordNet, and several other data sets.

The recent LINDA algorithm [7] shows how such links between identifiers can be discovered automatically in a scalable way. The algorithm was designed for the Hadoop distributed computing platform, which means that even very large crawls of the Web of Linked Data with billions of triples can be supplied as input. A new data set can therefore automatically be linked to many other data sets.

In conclusion, there are both incentives and tools for us to connect the data sets we build

and use. As a community, we should seek to identify and support identifier schemes that can serve as de facto standards. Publishers of linguistic data are strongly encouraged to link their resources to other existing data sets, e.g. in the rapidly growing cloud of Linguistic Linked Data. These efforts will lead to a much more useful Web of Linked Data.

3.9 Abstract

Thierry Declerck (DFKI Saarbrücken)

1. **What are in your view the most important challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web?**
   There is a (correct) statement that most knowledge is conveyed by Human Language, and therefore a criticism that you find in the Semantic web (I consider here mainly the LOD/LD instantiation of the Semantic web) only structured abstract knowledge representation. As a response to this criticism, our work can stress that language processing has to structure language data too, and that one of our task would be to represent structured language data in the same way as the knowledge objects, and to interlink those in a more efficient way as this has been done in the past, like for example in the simple/parole or the generative lexicon, linking thus language data "in use" with knowledge data "in use".
2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?**
   The possible approach sketched under point 1) would be deployed in a multilingual fashion. If multilingual data is successfully attached to knowledge objects, then multilingual and cross-lingual retrieval of knowledge is getting feasible. Not on the base of machine translation (only), but rather on the base of multilingual equivalents found linked to knowledge objects. At the end not only knowledge of the world can be retrieved, but also knowledge of the words (or language) associated with the knowledge of the world. The knowledge of the language would be partial (no full grammar is to be expected), but it can serve in many applications.
3. **Which figures are suited to quantify the magnitude or severity of the problem?**
   I can not answer concretely this question. I also do not know if there is a real "problem". We could go on the way we are doing by now (searching Google or the like, using domain specific repositories, using Question/Answering systems for accessing knowledge in text, etc), but I expect a gain of efficiency in many natural language based application, dealing with the treatment of knowledge: semantic annotation, semantic disambiguation, information extraction, summarization, all in multi- and cross-lingual contexts. Terminology should also benefit from this approach (linking multilingual linguistic linked data with linked data), in offering a better harmonization of the domain specific terms used in various languages, while referring to established terms used in the LD/LOD.
4. **Why do current solutions fail short?**
   Well: all the natural language expressions available in knowledge objects are not (yet) available in a structured form, reflecting the knowledge of language. So that the linking

of conceptual knowledge and language is done on a non-equilibrated manner: structured data on the one side and analysed strings on the other one.

5. **What insights do we need in order to reach a principled solution? What could a principled solution look like?**
   See the comment under point 1).
6. **How can standardization (e.g. by the W3C) contribute?**
   Giving an consensual view on representation of the various types of knowledge and ways to integrate those, by merging (OWL?) or by mapping/ linking (SKOS, lemon-LMF).

My possible contribution to the Workshop: Describing the potential LabelNet that can be resulting on the generalisation of linking structured language knowledge with domain knowledge. Generalizing the use of certain words/ expressions (phrases, clauses, etc) so that labels (or linguistically described terms) can be re-used in different knowledge contexts. There is also a specific domain I am working on, besides finance (XBRL; MFO) and radiology (RADLEX): Digital Humanities, more specifically two classification systems for tales and related genres. I am using there the Thompsom Motif Index and the Aarne Thompson Uther Type index of tales and transformed those in explicit taxonomies. We are also currently working on representing the labels of such taxonomies in LMF/lemon. I could present actual work in any of these 3 domains, if wished.

3.10 Supporting Collaboration on the Multilingual Semantic Web

Bo Fu (University of Victoria, British Columbia, Canada)

In relation to realising cross-lingual data access on the multilingual semantic web, particularly through the use of mappings, a lack of collaboration support in the current research field appears to be an important problem that is yet to be addressed.

One of the best examples of collaboration on the web during the past decade is Wikipedia, which has successfully demonstrated the value and importance of collaboratively building domain knowledge in a wide range of subject matters. Similarly, on the semantic web, knowledge bases (i.e. ontologies and other formal specifications) regardless of their representations or syntaxes, are the wisdom of communities and likely to involve the effort of individuals and groups from many different backgrounds. Given these characteristics, it is thus important to provide the necessary support for collaborations that are taking place during various stages on the semantic web.

In recent years, we have seen ontology editors integrating collaboration features. For instance, WebProtégé [76] is designed to support the collaborative ontology editing process by providing an online environment for users to edit, discuss and annotate ontologies. This trend in providing collaboration support is not yet evident in other semantic web research fields. For example, research in ontology mapping generation and evaluation has focused on developing and improving algorithms to date, where little attention has been placed on supporting collaborative creation and evaluation of mappings.

Proceeding forward, one of the challenges for the multilingual semantic web is to design and develop collaboration features for tools and services, in order for them to

- support social interactions around the data [28], so that a group of collaborators working on the same dataset can provide commentary and discuss relevant implications on common ground;
- engage a wider audience and provide support for users to share and publish their findings, so that information is appropriately distributed for group decision making;
- support long-term use by people with distinct backgrounds and different goals, so that personal preferences can be fully elaborated; and
- enhance decision making by providing collaborative support from the beginning of the design process, so that collaborative features are included in the design process of tools and services to prevent these features being developed as an afterthought.

3.11 Cross-lingual ontology matching as a challenge for the Multilingual Semantic Webmasters

Jorge Gracia (Universidad Politécnica de Madrid)

Recently, the Semantic Web has experienced significant advancements in standards and techniques, as well as in the amount of semantic information available online. Nevertheless, mechanisms are still needed to automatically reconcile information when it is expressed in different natural languages on the Web of Data, in order to improve the access to semantic information across language barriers. In this context several challenges arise [34], such as: (i) ontology translation/localization, (ii) cross-lingual ontology mappings, (iii) representation of multilingual lexical information, and (iv) cross-lingual access and querying of linked data. In the following we will focus on the second challenge, which is the necessity of establishing, representing and storing cross-lingual links among semantic information on the Web. In fact, in a "truly" multilingual Semantic Web, semantic data with lexical representations in one natural language would be mapped to equivalent or related information in other languages, thus making navigation across multilingual information possible for software agents.

Dimensions of the problem

The issue of cross-lingual ontology matching can be explored across several dimensions

1. Cross-lingual mappings can be established at different knowledge representation levels, each of them requiring their own mapping discovery/ representation methods and techniques: i. conceptual level (links are established between ontology entities at the schema level), ii. instance level (links are established between data underlying ontologies), and iii. linguistic level (links are established between lexical representations of ontology concepts or instances).
2. Cross-lingual mappings can be discovered runtime/offline. Owing to the growing size and dynamic nature of the Web, it is unrealistic to conceive a Semantic Web in which all possible cross-lingual mappings are established beforehand. Thus, scalable techniques to dynamically discover cross-lingual mappings on demand of semantic applications have to

[28] In this context, data can be any variable related to applications on the semantic web. For example, it can be the results from ontology localisation, ontology mapping or the evaluations of such results.

be investigated. Nevertheless, one can imagine some application scenarios (in restricted domains for a restricted number of languages) in which computation and storage of mappings for later reuse is a viable option. In that case, suitable ways of storing and representing cross-lingual mappings become crucial. Also mappings computed runtime could be stored and made available online, thus configuring a sort of pool of cross-lingual mappings that grows with time. Such online mappings should follow the linked data principles to favour their later access and reuse by other applications.

3. Cross-lingual links can be discovered either by projecting the lexical content of the mapped ontologies into a common language (either one of the languages of the aligned ontologies or a pivot language) e.g., using machine translation, or by comparing the different languages directly by means of cross-lingual semantic measures (e.g., cross-lingual explicit semantic analysis [74]). Both avenues have to be explored, compared, and possibly combined.

What is needed?

In summary, research has to be done in different aspects:

- *Cross-lingual ontology matching.* Current ontology matching techniques could be extended with multilingual capabilities, and novel techniques should be investigated as well.
- *Multilingual semantic measures.* Such novel cross-lingual ontology matching techniques above mentioned have to be grounded on measures capable of evaluating similarity or relatedness between (ontology) entities documented in different natural languages.
- *Scalability of matching techniques.* Although the scalability requirement is not inherent to the multilingual dimension in ontology matching, multilingualism exacerbates the problem due to the introduction of a higher heterogeneity degree and the possible explosion of compared language pairs.
- *Cross-lingual mapping representation.* Do current techniques for representing lexical content and ontology alignments suffice to cover multilingualism? Novel ontology lexica representation models [55] have to be explored for this task.

3.12 Abstract

Iryna Gurevych (Technical University Darmstadt)

We first outline a set of research directions for the multilingual content processing on the web, such as aggregating the knowledge in multiple documents, assessing the quality of information, engineering complex multilingual Web-based systems, and scalability of machine learning based approaches to new tasks and domains. Then, we present some research initiatives at UKP Lab with immediate relevance to the research directions listed above.

Research directions

The volume of text-based data, especially user-generated content in many languages, on the Web has been continuously growing. Typically, there are multiple documents of various origins describing individual facets of the same event. This entails redundancy, resulting in the need to aggregate the knowledge distributed across multiple documents. It involves the tasks such as removing redundancy, information extraction, information fusion and text

summarization. Thereby, the intention of the user and the current interaction context play an important role.

Another fundamental issue in the Web is assessing the quality of information. The vast portion of the content is user-generated and is thus not subject to editorial control. Therefore, judging its quality and credibility is an essential task. In this area, text classification methods have been applied and combined with social media analysis. Since the information on the Web might quickly become outdated, advanced inference techniques should be put to use in order to detect outdated content and controversial statements found in the documents.

Due to advances in ubiquitous computing and the penetration of small computer devices in everyday life, the integration of multiple knowledge processing techniques operating across different modalities and different languages on huge amounts of data has become an important issue. This is an issue with challenges to be addressed in software engineering. It requires standardization of the interface specifications regarding individual components, ensuring the scalability of approaches to large volumes of data, large user populations and real-time processing, and solutions regarding the technical integration of multiple components into complex systems.

Current multilingual language processing systems extensively utilize machine learning. However, the training data is lacking in many tasks and domains. To alleviate this problem, the use of semi- supervised and unsupervised techniques is an important research direction. For the supervised settings, utilizing crowdsourcing and human computation such as Amazon Mechanical Turk, Games with a Purpose, or Wiki-based platforms for knowledge acquisition is a current research direction [36]. Research is needed to find ways of efficiently acquiring the needed high-quality training data under the time and budget constraints depending on the properties of the task.

Research Initiatives at UKP Lab

The above research directions have been addressed in several projects by UKP Lab at the Technical University Darmstadt, described below.

Sense-linked lexical-semantic resources. We present a freely available standardized large-scale lexical-semantic resource for multiple languages called UBY[29] [26, 35]. UBY currently combines collaboratively constructed and expert-constructed resources for English and German. It is modeled according to the ISO standard Lexical Markup Framework (LMF). UBY contains standardized versions of WordNet, GermaNet, FrameNet, VerbNet, Wikipedia, Wiktionary and OmegaWiki. A subset of the resources in UBY is linked at the word sense level, yielding so-called mono- and cross-lingual sense alignments between resources [25, 58, 65]. The UBY database can be accessed by means of a Java-based API available at Google Code [30] and used for knowledge-rich language processing, such as word sense disambiguation.

Multilingual processing based on the Unstructured Information Management Architecture (UIMA). We put a strong focus on component-based language processing (NLP) systems. The resulting body of software is called the Darmstadt Knowledge Processing Software Repository (DKPro) [24]. Parts of DKPro have already been released to the public as open source products, e.g.:

[29] http://www.ukp.tu-darmstadt.de/uby
[30] http://code.google.com/p/uby

- *DKPro Core* [31] is an integration framework for basic linguistic preprocessing. It wraps a number of NLP tools and makes them usable via a common API based on the Apache UIMA framework. From the user perspective, the aim of DKPro Core is to provide a set of components that work off-the-shelf, but it also provides parameter setting options for the wrapped tools. The roadmap for DKPro Core includes: packing models for the different tools (parser, tagger, etc.) so they can be logically addressed by name and version and downloaded automatically, cover more tagsets and languages, logically address corpora and resources by name and version and download them automatically, provide transparent access to the Hadoop HDFS so that experiments can be deployed on a Hadoop Cluster.
- *DKPro Lab* [32] is a framework to model parameter-sweeping experiments as well as experiments that require complex setups which cannot be modeled as a single UIMA pipeline. The framework is lightweight, provides support for declaratively setting up experiments, and integrates seamlessly with Java-based development environments. To reduce the computational effort of running an experiment with many different parameter settings, the framework uses dataflow dependency information to maintain and reuse intermediate results. DKPro Lab structures the experimental setup with three main goals: facilitating reproducible experiments, structuring experiments for better understandability, structuring experiments into a workflow that can potentially be mapped to a cluster environment. In particular, the latter is currently in the focus of our attention.

The DKPro software collection has been employed in many NLP projects. It yielded excellent performance in a series of recent language processing shared tasks and evaluations, such as:

- Wikipedia Quality Flaw Prediction Task in the PAN Lab at CLEF 2012. [30]
- Semantic Textual Similarity Task for SemEval-2012, held at *SEM (the First Joint Conference on Lexical and Computational Semantics). [8]
- Cross-lingual Link Discovery Task (CrossLink) at the 9th NTCIR Workshop (NTCIR-9), Japan. [51]

3.13 Collaborative Community Processes in the MSW area

Sebastian Hellmann (Leipzig University)

This presentation introduces three major data pools that have recently been made freely available as Linked Data by a collaborative community process: (1) the DBpedia Internationalization committee is concerned with the extraction of RDF from the language-specific Wikipedia editions; (2) the creation of a configurable extractor based on DBpedia which is able to extract information from all languages of Wiktionary with manageable effort; (3) the Working Group for Open Lingustic Data, an Open Knowledge Foundation group with the goal of converting Open Linguistics data sets to RDF and interlinking them. The presentation highlights and stresses the role of Open Licenses and RDF for the sustenance of such pools. It also provides a short update on the recent progress of NIF (Natural Language

[31] http://code.google.com/p/dkpro-core-asl/
[32] http://code.google.com/p/dkpro-lab/

Processing Interchange Format) by the LOD2-EU project. NIF 2.0 will have many new features, including interoperability with the above-mentioned data pools as well as major RDF vocabularies such as OLiA, Lemon, and NERD. Furthermore, NIF can be used as an exchange language for Web annotation tools such as AnnotateIt as it uses robust Linked Data aware identifiers for Website annotation.

3.14 Overcoming Linguistic Barriers to the Multilingual Semantic Web

Graeme Hirst (University of Toronto)

Sometime between the publication of the original Semantic Web paper by Berners-Lee, Hendler, and Lassila [5] and Berners-Lee's "Linked Data" talk at TED[33], the vision of the Semantic Web contracted considerably. Originally, the vision was about "information"; now it is only about data. The difference is fundamental. Data has an inherent semantic structure and an *a priori* interpretation. Other kinds of information need not. In particular, information in linguistic form gains an interpretation only in context, and only for a specific reader or community of readers.

I do not mean to criticize the idea of restricting our Semantic Web efforts to data *pro tem*. It is still an extremely challenging problem, and the results will still be of enormous utility. At the same time, however, we need to keep sight of the broader goal, and we need to make sure that our efforts to solve the smaller problem are not just climbing trees to reach the moon.

In the original vision, "information is given well-defined meaning" (p. 37), implying that it didn't have "well-defined meaning" already. Of course, the phrase "well-defined meaning" lacks well-defined meaning, but Berners-Lee et al. are not saying that information on the non-Semantic Web is meaningless; rather what they want is precision and lack of ambiguity in the Semantic layer. In the case of linguistic information, this implies semantic interpretation into a symbolic knowledge representation language of the kind they talk about, which is a goal that exercised, and ultimately defeated, research in artificial intelligence and natural language understanding from the 1970s through to the mid-1990s.

One of the barriers that this earlier work ran into was the fact that traditional symbolic knowledge representations – the kind that we still see for the Semantic Web – proved to be poor representations for linguistic meaning, and hierarchical ontologies proved to be poor representations for the lexicon [39]. Near-synonyms, for example, form clusters of related and overlapping meanings that do not admit a hierarchical differentiation. And quite apart from lexical issues, any system for representing linguistic information must have the expressive power of natural language; we have nothing anywhere close to this as yet.

All these problems are compounded when we add multilinguality as an element. For example, different languages will often present a different and mutually incompatible set of word senses, as each language lexicalizes somewhat different categorizations or perspectives of the world, and each language has lexical gaps relative to other languages and to the categories of a complete ontology. It is rare even for words that are regarded as translation

[33] The Next Web, TED Conference, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

equivalents to be completely identical in sense; more usually, they are merely cross-lingual near-synonyms [39].

And then we have the problem of querying linguistic information on the Semantic Web, again in a natural language. Much of the potential value of querying the Semantic Web is that the system may act on behalf of the user, finding relevance in, or connections between, texts that goes beyond anything the original authors of those texts intended. That is, it could take a reader-based view of meaning, "What does this text mean to me?" [38]. The present construal of the Semantic Web, however, is limited to a writer-based view of meaning. That is, semantic mark-up is assumed to occur at page-creation time, either automatically or semi-automatically with the assistance of the author [5]; a page has a single, fixed, semantic representation that (presumably) reflects its author's personal and linguistic worldview and which therefore does not necessarily connect well with queries to which the text is potentially relevant.

But that's not to say that author-based mark-up isn't valuable, as many kinds of natural language queries take the form of intelligence gathering, "What are they trying to tell me?" [38]. Rather, we need to understand its limitations, just as we understand that the query "Did other people like this movie?" is an imperfect proxy for our real question, "Will I like this movie?".

This gives us a starting point for thinking about next steps for a monolingual or multilingual Semantic Web that includes linguistic information. We must accept that it will be limited, at least *pro tem*, to a static, writer-based view of meaning. Also, any semantic representation of text will be only partial, and will be concentrated on facets of the text for which a representation can be constructed that meets Berners-Lee et al.'s criterion of relative precision and lack of ambiguity, and for which some relatively language-independent ontological grounding has been defined. Hence, the representation of a text may be incomplete, patchy, and heterogeneous, with different levels of analysis in different places [40].

We need to recognize that computational linguistics and natural language processing have been enormously successful since giving up the goal of high-quality knowledge-based semantic interpretation 20 years ago. Imperfect methods based on statistics and machine learning frequently have great utility. Thus there needs to be space in the multilingual Semantic Web for these kinds of methods and the textual representations that they imply – for example, some kind of standardized lexical or ontolexical vector representation. We should expect to see symbolic representations of textual data increasingly pushed to one side as cross-lingual methods are developed in distributional semantics [29] and semantic relatedness. These representations don't meet the "well-defined meaning" criterion of being overtly precise and unambiguous, and yet they are the representations most likely to be at the centre of the future multilingual Semantic Web.

3.15 Interoperability in the MSW

Chu-Ren Huang (The Hong Kong Polytechnic University)

1. The most crucial challenge to a multilingual semantic web is its accessibility and interoperability for people from different linguistic and cultural backgrounds, a challenge that the currently envisioned shared ontology could compound rather than ameliorate.

First, the semantic web and its content must be accessible to people from different parts of the world using different languages and having different culturally conventionalized world views. This issue requires both multilingual language resources and culture-specific ontology, presumably linked through and mapped to a shared ontology. Second, I consider the inter-operability issue crucial but from a maverick perspective. It is crucial to recognize that for all the tasks performed on the semantic web, each of them comes with an intentional goal with culture-specific background. Taking Tim Berners-Lee's example of buying flowers, the typical and most likely event type in the West is to buy loose flowers or bouquets for someone dear to the buyer; but in the Chinese context, the typical event is to buy flower basket installations for social networking functions such as a wedding or opening of a business. Similarly, when searching for a family diner out, a diner or a "family-oriented" restaurant (such as Appleby's) with crayons for children are typical for users in the U.S. But in the Chinese context, a Chinese meal with round-table seating may be crucial. I view inter-operability challenge as 1) to be able to identify these common event types as well as their event structure skeletons and cultural variants for integration of information; and 2) to allow culture/domain specific event-driven tasks to exploit knowledge content encoded in either the shared ontology or a domain specific ontology. It is of course important to note that these event-variation issues are often embedded in language and need to be described in languages accessible to the users.

2. Accessibility and Interoperability as described above is critical to whether an industry based on SW can deliver or not. In the multilingual and multicultural world connected by the SW, localization will not be as effective as SW is concept-based, not text-based; and in our increasingly multi-cultural world, a user's assumed cultural convention is rarely simple and very often not determined by the language s/he uses (or his/her IP).
   All sectors should be affected. However, this challenge should be particularly keen for the following sectors, (1) creative culture industries (CCI), including but not limited to (culture) tourism, hospitality, (digital) museums, etc., (2) health communication and health care information providers, (3) advertisement. (Second point (2) is skipped as no relevant data can be given, underlining how difficult it is to characterize the cultural/linguistic background of web users.)
3. It is likely that the current solutions fail short because they focus on ensuring the meaning content is accessible to different machines, but not to how the information can be utilized or interpretable by human users in the world. It is also important to note that
   - there are no (or at least very rare) large scale culturally sensitive knowledge bases;
   - construction of ontologies, including domain-specific ontologies, so far focused on shared, not differential knowledge structure.
4. People often act on both personal experience and culturally conventionalized shared experience. Note these experiences can correspond to shared knowledge, but do not necessarily follow the knowledge structure represented in an upper ontology. In addition, such behaviors are driven by the goal of the person (i.e. the telic cause of Aristotle).
   Take the treatment of environmental and ecology issues for example. It seems that all issues with environmental impact can be boiled down to those which impacts feeding or breeding for the living organisms involved. However, feeding for person, for a cow, or for a microbe, involves very different participants and very different environmental conditions; the current approach to SW ontologies seems to require that these events are given radically different representations. Another good example involves emotions. Although there are culture- and species-specific variations to expressing and reaction to emotions, it is generally accepted that people recognize the same emotion types across

different cultures: anger, fear, etc. The recognition of these common event types (e.g. feeding, breeding, fear, happiness, etc.) given different contextual information will endow SW with powerful and effective means to deliver what users really want.
The solution is an additional dimension of ontology based on event types in addition to the current entity type based shared ontology. The Chinese writing system as an ontology is a good example, as I have shown previously that the conceptual cluster sharing a radical is often based on event-relations such as telic and agentive, and less often by entity-type relations such as is-a or part-of.

5. Standardization should help, provided that we do thorough studies to explore the common event-types which are crucial to human activities and draft event-driven ontologies based on the research. An especially difficult challenge is to capture the telic event types, being able to link telic goals to events that are necessary to attend that goal will allow SW to work on both meaning stated and intension/need expressed.

3.16 Rendering lexical and other resources as Linked Data

Nancy Ide (Vassar College, New York, USA)

A "language- and culture-neutral" Semantic Web (SW) will have to accommodate different linguistic and cultural perspectives on the knowledge it contains, in the same way as it must accommodate temporal, spatial, etc. perspectives. In the long term, probably the greatest challenge for SW development is to seamlessly handle a multiplicity of viewpoints (including language) on knowledge.

In the short term, we can do what we can with what we have now. One hypothesis is that a multilingual SW can be achieved – or at least approached – by mapping other languages to the broad range of existing ontological vocabularies that have been developed, almost exclusively in English, for various topics and domains. Existing language resources that cover multiple languages – notably, resources such as WordNet and FrameNet, but also bi- and multi-lingual lexicons developed in, for example, large EU projects over the past two decades – could be exploited for this purpose.

A major step toward a multilingual SW, and toward interoperability for mono- and multi-lingual language resources in general, would be to render lex- ical and other language resources as linked data (as WordNet and FrameNet already are). As linked data, the resources will have achieved some degree of structural ("syntactic") interoperability (to the extent that the relations and properties used in their representations are defined consistent). A linked data representation will also make a move toward conceptual ("semantic") interoperability (see [43]), because the various resources can, in principle, be linked to each other, either directly or via mediator ontologies that provide an exchange reference model for linking between resources (see, for example, Chiarcos' Ontologies of Linguistic Annotation (OLiA) [14]). While mapping concepts among lexicons and similar resources is notably difficult, some immediate steps can be made by linking the linked lexicons with corpora annotated with the categories defined in them. For example, the Manually Annotated Sub-Corpus (MASC) [42] has been rendered in RDF, and its WordNet and FrameNet annotations have been linked to the relevant entries in the linked data versions of these resources [15]. In this form, all three resources are automatically combined, and SPARQL queries can be used to access, for

example, all occurrences annotated with a specific FrameNet frame element, or all words used in a particular WordNet sense. Additionally, one could query to find the FrameNet frames that correspond to a particular WordNet sense via the corpus annotation, thus providing input for a better conceptual mapping of the two resources themselves. Ultimately, any annotated word, phrase, entity, etc. could be linked to occurrences of the same phenomenon in data in other languages, either directly or via an interlingua (if appropriate). The potential of such massively interlinked multilingual data for NLP research and development is enormous.

Rendering language resources as linked data, so that they can be used together seamlessly, requires a consistent model of the phenomena in question, including not only ontological concepts, but also their inter-relations and properties. There have been efforts to devise standards which, although not specifically aimed toward linked data, provide underlying models of lexical data [31] that are isomorphic to the linked data model. However, there is much more (relatively mundane) work to be done in order to ensure compatibility in the domain model. As a simple example, consider modeling the general concept "Annotation" by defining its relations to other concepts and properties, with an eye toward enabling relevant queries. This requires answering (seemingly) simple questions like, does an Annotation have a single target, or can it have several? If it has several, does it apply to each individually (e.g., a "noun" annotation applied to all text spans identified as nouns throughout a text), or does it apply to an aggregation of its targets (e.g., a "verb" annotation applied to two discontiguous text spans that comprise a single verb)? How do we distinguish these two cases in the model? etc. While this seems trivial on the one hand, the different communities for whom "annotation" is a relevant concept, including not only computational linguists but also humanities scholars who annotate in the more traditional sense, as well as annotators of images, video and audio, etc. – must adopt a common, or at least compatible, model if their data is to be used together (which much of it will ultimately be), and if we do not want to be faced with a different form of query for each case. This is where standards-making bodies, like W3C and ISO, must critically step in.

Groups such as the W3C and ISO can foster the development of a (multi- lingual) SW by:

- Promoting the use of SW technologies to structure and describe existing and future language resources, including lexicons, ontologies, and corpora developed by the NLP community;
- Establishing stronger ties with the NLP community, to leverage their expertise in identifying/extracting information;
- Continuing the top-down development of the SW infrastructure;
- Developing best practice guidelines for domain modeling in RDF/OWL;
- Overseeing and coordinating bottom-up development of RDF/OWL mod- els for specific bodies of knowledge developed by specific disciplines/ communities/ interests;
- Actively seeking to identify commonalities among the varied models and bodies of knowledge and ensuring that efforts are combined/harmonized;
- Working on ways to accommodate different views of knowledge, including language and culture, in the SW.

Comprehensive interoperability via standardization is a long-term goal that is unlikely to be achieved anytime soon. This means that for the interim, we have to explore effective ways to bridge the differences in the concepts and structure of knowledge sources. Representing existing language resources as linked data is one way to approach that problem.

3.17 Leveraging MSW research for practical applications: what can we do?

Antoine Isaac (Europeana)

Caveat: the views here represent a personal take on MSW issues, resulting from involvement in Europeana.eu (providing access to 24 millions objects from 33 countries) and other projects in the cultural sector, as well as my experience with SKOS and ontology alignment. It shall be taken with a grain of salt: multilingual issues have not been the exclusive focus of my work and I may miss some efforts, especially recent ones. Which in this case tells about their visibility.

Technical & practices issues.

SW technology does a lot to enable ML: tags on RDF literals allow for language-specific data, and the aggregation of those language-specific literals allow for "multilingual entities". Some ontologies enable easy representation of e.g., multilingual concepts (SKOS). There are even finer-grained models available (Lexinfo). But they are seldom known and used in large datasets. As a result some technical issues still don't have a commonly shared solution. For example, representing "translation of a statement". Consider the following:

```
ex:book   dc:subject   "multilingual semantic web"@en ;
          dc:subject   "challenges"@en ;
          dc:subject   "web sémantique multilingue"@fr .
```

This is a typical example of bibliographic data ported to RDF in a very basic way. It does not fully represent translation links and thus fails to some applications. There should also be more attention given to patterns for giving the language of an object (e.g. a web page or a video recording) vs. the language of data about that object or the language of an interface. Technology (e.g. through the "one-to-one principle") makes clear what can be done; but in some domains (Europeana) data providers or consumers may still be confused.

Availability of tools and data

Many tools of reference in the SW community are mostly monolingual [34]. A lot of datasets, and most experimentations and case studies are in English. More precisely, there are many resources available for a multilingual SW:

- terminological/conceptual bases (wordnets, SKOS datasets, dictionaries...)
- language processing tools (translators, language recognizers, parsers...)

Yet these are difficult to find. Few inventories exist, mostly paper reports not easily exploitable. We need a better way of gathering and sharing information on MSW resources. Here, specific metrics can be useful, for evaluating multilingual tools and measuring the "multilingual quality" of datasets. One starting point would be to indicate the language(s) covered by datasets on the linked data cloud, the labels per language, etc., refining for example language-related quality criteria used in the SKOS community (e.g., [54]).

[34] Marlies Olensky. Market study on technical options for semantic feature extraction. Europeana v2.0 technical report, http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Market+study+on+technical+options+for+semantic+feature+extraction

Further, many relevant resources are not open and/or are published in a format that does not enable easy re-use. This the case for many wordnets, and a true pity for resources that are created with public money. Some communities have made progress in releasing multilingual datasets, but a lot remains to be done.

In relation with the above metrics, guidelines could help (if not standards) both to provide MSW resources or to select them for consumption.

To compensate the rarity of tools and resource, we should also be open to "less AI-focused" solutions, such as crowdsourcing translation or multilingual tagging.

Community organization and awareness

The above issues are partly caused by the homogeneity of the "core" SW community, mostly academic and English-speaking. That prevents diversity to emerge re. experiments and tools. It also makes it harder to be aware of relevant efforts in other communities (information retrieval, databases, more general web community); if just because these communities have the same bias...

Further, the difficulty of "traditional" problems is raised an order of magnitude higher when transferred from a monolingual context to a multilingual one (NB: that applies both re. finding and implementing solutions, and evaluating them). Bluntly put: working on a multilingual problem is not the most effective way of getting a paper published, and that does not help. For example the Ontology Alignment Evaluation Initiative (OAEI[35]) has featured multilingual tracks for a while. Bar a few exceptions [57], participation has often been low.

Maybe the current SW community is not the ideal forum for tackling multilingual problems. Or it may just be able to progress on very specific issues (e.g. focusing on producing and sharing data). On evaluation matters, especially it could help to better share efforts (corpora, gold standards, methods and measures) with other communities – e.g., databases, web (services) or information retrieval. A relevant initiative is CLEF [36].

Besides, there are many vendors that propose relevant solutions, especially in "language technology". But they probably find it difficult to relate to the SW community. As long as vendors make a reasonable benefit in their current (non-SW) environment many won't seriously move and liaise with academic efforts. We need to getting more varied people interested and contributing to the MSW issues – but maybe from their own communities' perspectives.

Use cases

For bringing people together, it would help to identify the most relevant features of (end user) application scenarios. One usual example is localization: adapt a display depending on the language/country selected. This imposes multilingual requirements both on ontologies and instance data level. But there are other dimensions to multilingual access in which semantic web technology can be relevant: query translation, document translation, information extraction and data enrichment, browsing and personalization, knowledge acquisition for non English speakers, interaction between users and system... Some of these are neither strictly multilingual nor semantic web-specific. In such case, the potential added value of

[35] http://oaei.ontologymatching.org/

[36] Conference and Labs of the Evaluation Forum, formerly known as Cross- Language Evaluation Forum. http://www.clef-initiative.eu/

MSW should be detailed: for example, enhancing search results in one language based on links established using semantic resources in another language.

Maybe such a gathering focus more on cases where multilinguality is really crucial. For example, Europeana is encouraging application of SW technology for access to culture resources, where all EU languages should ultimately been tackled. It especially envisions tapping into a semantic data layer, which involves alignment of multilingual vocabularies and metadata enrichment. In a completely different domain, the VOICES project envisions using linked data technology for social and rural development. Key issues there are sharing locally produced data where local languages are more important than English, and building a robust data-to-speech service [18].

3.18 Practical challenges for the multilingual Semantic Web

Christian Lieske (SAP AG)

When people start to think about cultural diversity, they sooner or later start to talk about mouth-watering dishes. Thus, I have taken the freedom to choose a title for my position statement that alludes to a recipe (see below). As in almost any recipe, important details are missing from my recipe – the details can only be revealed face-to-face/in a joint session of practice. Accordingly, I would be happy to be invited to elaborate my position/my input at the seminar.

Rather than providing direct references, I provide pointers that indicate with which kind of background I tend to look at things.

Please note: All my thoughts are my own and not endorsed by my employer in any way.

10 Ingredients for the Multilingual Semantic Web Delight

1. *World-ready Web Stack:* Look for example at the mailing lists run by the W3C Internationalization Activity, or the Unicode Consortium to see that even mainstream topics such as HTML5, CSS3, JavaScript and Unicode are still undergoing modification in order to cover multilingual/- cultural dimensions more complete, or enhanced.
2. *Concept-based Content Creation:* Realize how model-based approaches (doesn't matter if your model entities are objects or events) already help to generate expressions in multiple idioms. In addition, read up for example on the Wikidata project to sense that there is a value in language-neutral representations.
3. *Connected Organizational Constituencies:* Be surprised that the problem space of that EC project epSOS is overlapping with the "Spanish patient needs medication in Germany" scenario that shows up in Semantic Web scenarios.
4. *Transparency on Stakeholders/Contributors and Contribution Framework:* Don't be blind to the fact that enterprises may have much to contribute to the Semantic Web. Don't you think for example that providers of pharmaceutical companies already have databases that capture relationships between the incarnations of their products on different markets?
5. *Eye on Multi-Modality:* Acknowledge that human interaction and information dissemination is not just based on written text. Consider how to take care of graphics/images or sound/voice (especially considering findings on the situation in the non-Western world from initiatives such as the World Wide Web Foundation or the Praekelt Foundation).

6. *Anyone-Anytime-Anywhere Paradigm:* Provide tooling that doesn't require a diploma in SPARQL, and the installation of a heavyweight application. If you want for example contributions to ontologies or vocabularies think "Point your mobile browser to a URL, and comment". Understand for example tools like translation memories engines that assist in language-related activities.
7. *Reuse/Minimalism, and Clean, Open, Traceable Information Sources:* Ask yourself how much trust you would have in a HTML table that would tell you "The drug that is called X in Spain is called Y Germany". Wouldn't you for example like to see provenance information before you order the drug?
8. *Open-Minded NLP Community:* Be aware of the fact that "mapping" is a very simple mathematical function. Do we think for example that a mapping will suffice to go from a blood pressure as measured in Germany to one that can be understood by a French speaking physician? Or does the "mapping" concept that is favoured by the NLP community need to be rethought?
9. *Non-functional Requirements:* Don't underestimate that you may need to prioritize, and schedule roadmap items. Otherwise, decision makers may not see the relevance and importance of Semantic Web activities.
10. *Implementation-backed Standards and Best Practices:* Makes sure that you have implementations for standards and best practices. Think for example how much easier the creation of multilingual Web sites would be if all Web server downloads would come with a "template" for multilanguage/ multi-country Web presences.

3.19 Localization in the SW: the status quo

John McCrae (Bielefeld University)

While some of the key resources in the Semantic Web, notably DBpedia, have placed considerable effort on internationalisation of their resources, most of the vocabularies including some of the widely used vocabularies, such as FOAF and even W3C standards such as RDFS, fail to provide labels in any language other than English. Even worse, many of these resources fail to even indicate that these labels are in English by means of standard meta-data. A clear issues with providing multilingual data as labels is that, all the emphasis on providing labels in languages other than English is on the data provider. It is of course extremely unlikely that for all but the largest of data providers, they could provide and check translations for even the Top 10 languages[37]. Even worse, for many applications that wish to use linked data and the Semantic Web, more lexical information is required than just a simple label. In particular, for many applications, such as question answering and natural language generation[38], information such as part-of-speech, morphology (e.g., irregular inflections) and syntactic information, such as subcategorization, would be extremely helpful. Anomalously, a simple and clear solution to this is available: by linking to dictionaries and lexica we can clearly define these concepts along with their multilingual equivalents.

[37] The top 10 languages by average ranking in GDP and number of speakers are: English, Chinese, Spanish, Japanese, Arabic, German, Portuguese, Hindi, Russian and French
[38] Such methods are required by intelligent personal assistants such as Apple's Siri.

A recent paper by Basil Ell et al. [27] recently claimed that only 0.7% of the entities in the web of data had labels in a language other than English, while 38.3% of the data had English labels. As such it is clear that the adoption of multilingual linked data within industry and research has been severely limited. Assuming that these organisations do not have the resources to provide translations for most language, a key issue is how these translation may be sourced from third parties. A solution to this may be to provide a central repository for localisation of vocabularies and data, i.e., a Google or Facebook of multilingual data. While this solution may have many clear advantages it seems unlikely that any existing service provider would step up to fill this role nor that it would a profitable new venture. As such, it seems that this solution is unlikely to materialise soon and instead linking to dictionaries seems to be the more feasible solution, and has the advantage that the creation of multilingual lexical data is now performed by those who have the interest and knowledge in doing so, instead of being a requirement on all data providers.

Based on the assumption that we need to link to entities defined in dictionaries and lexica, it is clear that there is some need for standardisation to define how this linking is performed and more importantly what format should be expected at when dereferencing such a link. This could happen in likely two ways: either the creation of a single large data source, likely based on a community based Wiki interface, causing a de facto standardisation, or preferably a standard format, introduced by organisations such as W3C or ISO, that allows for a competitive but inter-operable ecosystem for the description of such multilingual data. As such we [55] have proposed such a model we call lemon, the "Lexicon Model for Ontologies", that aims to allows ontologies and linked data in existing semantic formats such as OWL and RDFS to be linked to rich lexical descriptions. We have continued to develop this as part of the OntoLex community group[39], with the aim of creating a linguistically sound model that will provide a guiding paradigm for producers of linked data lexica.

3.20 Localization and interlinking of Semantic Web resources

Elena Montiel Ponsoda and Guadalupe Aguado de Cea (Universidad Politécnica de Madrid)

Some of the most important challenges in providing multilingual access to the Semantic Web (SW) are related with two aspects:

1. The provision of multilingualism to ontologies and data sets documented in one natural language (NL)
2. The interlinking or mapping of semantic resources documented in different NLs

The interlinking or mapping of semantic resources documented in different NLs As the Open Linked Data phenomenon has shown, more and more resources are published in languages other than English [60]. A truly multilingual access to the SW involves, in our opinion, either the localization or translation of some resources to several NLs, or the establishment of links between and among ontologies and data sets in the same domain described in different NLs.

The main problem in the localization or interlinking matter is the fact that ontologies or data sets in the same domain may present some of these aspects:

[39] http://www.w3.org/community/ontolex

- conceptualization mismatches
- different levels of granularity
- different perspectives of the same domain

Some of these aspects are also present in the interlinking of resources available in the same language, or what is the same, in the interlinking or mapping of monolingual resources.

In the localization of resources, current solutions fall short because of several reasons:

- No homogeneous representation mechanisms accepted by the community are available. In this sense, several ontology-lexicon models proposed in the last years have tried to overcome this problem (LIR [59], lemon [56]).
- Solutions fall short of accounting for conceptualization mismatches

We argue that in the localization of ontologies, specific representation models have to be able to define specific relations between NL descriptions in different languages, what we term translation relations or cross-lingual relations.

Highly related with this issue is the representation of term variation at a monolingual or multilingual level. A term variant has been defined as ´´an utterance which is semantically and conceptually related to an original term" [16]. We believe that the representation of term variants would also contribute to the establishment of links or relations between the NL descriptions associated to concepts (within or across languages).

A further problem may involve the automation of the localization process (See some proposed approaches for the automatic localization of ontologies [28, 55]).

As for the interlinking of resources in different NLs, there are no specific links or mappings that can account for conceptualization mismatches between or among resources in several NLs. Some available solutions could be:

1. The "equivalence" or "sameAs" link would represent a solution in the case that highly similar conceptualizations are available for the same domain in different languages.
2. The "skos:broader" or "skos:narrower" link would work in some cases but their semantics are not clearly defined.

In both cases, the localization and the interlinking, in order to reach a principled solution we would need to provide well defined representation mechanisms and mappings intended principally to account for the differences between conceptualizations in different NLs. Without any doubt, standardization can play a key role to help solving this matter.

3.21 Multilingual Word Sense Disambiguation and the Semantic Web

Roberto Navigli (Sapienza University of Rome)

Motivation.

The Web is by far the largest information archive available worldwide. Seen as a vast repository of text, the Web contains the most disparate information which can virtually satisfy all the possible user needs. However, nowadays the textual information needed by a user, such as in news, commentaries and encyclopedic contents, is provided in an increasing number of languages. For example, even though English is still the majority language, the Chinese and Spanish languages are moving fast to capture their juicy Web share, and more

languages are about to join them in the run. The prototypical example of this trend is Wikipedia, whose multilingual growth is clearly exponential[40].

However, the Web suffers from two important issues:

1. First, the vast majority of textual content is not linked to existing ontologies, because of:
   - the paucity of ontologies for several domains,
   - the lack of a suitable lexicalization for the concepts within many existing ontologies.

   While much effort has been devoted to the creation of ontologized information, the current state of the art is still very far from tackling the lack of domain and lexical coverage.
2. Second, the truly multilingual nature of today's Web is currently a barrier for most users, rather than an opportunity for having more and richer information. For instance, recently Google announced that Google Translate features about 200 million users per month[41], many of which are using mobile devices to obtain the appropriate lexicalization of their information need in another language. This need is also testified by the yearly organization of cross-lingual Information Retrieval forums like CLEF[42].

These key issues cry for the need of frameworks, tools and algorithms aimed at addressing the interactions between ontological representations and a babel of languages, so as to provide smooth access to multilingual content on the Web. The beneficiaries of such seamless integration would not only be end users, but also SMEs in virtually all industry sectors whose business is connected to the Web. In fact, an infrastructure able to overcome the language barrier would open new business opportunities in any domain, by increasing the customer base and approaching markets in new countries and regions.

Today's research in brief.

The two main research communities concerned with the above-mentioned issues are, on the Web side, the Semantic Web (SW) community, and, on the language side, the Computational Linguistics (CL) community. On the one hand, the SW community has conducted much work in the direction of addressing the tasks of ontology construction, learning and population [9, 10], ontology linking [67], ontology matching [75], etc. On the other hand, the CL community has increasingly been working on important issues such as multilinguality [52, 63, 71, 72], disambiguation [61] and machine translation [53].

What comes next.

As suggested above there is an important missing link between the two communities, i.e., integrating ontologies with languages. Important efforts in this direction are DBPedia, YAGO, WikiNet, MENTA and Freebase. However, none of these proposals aims at bringing together the two worlds of the SW and CL by jointly and synergistically addressing the issues of ontological solidity, multilinguality and sense ambiguity. For instance, DBpedia is mainly concerned with popular types of Named Entities and manually maps concepts to WordNet, YAGO maps Wikipedia entities to the first senses of WordNet lemmas, MENTA addresses the multilinguality issue, focusing on the taxonomical aspect of an ontology, etc. In my group, an ambitious project – funded by the European Research Council – is currently under way, with two main goals: first, creating BabelNet [63, 64], a large, wide-coverage multilingual semantic

[40] http://meta.wikimedia.org/wiki/Wikimedia_in_figures_-_Wikipedia
[41] http://mashable.com/2012/04/26/google-translate-users/
[42] http://www.clef-initiative.eu/

network for dozens of languages and with many kinds of semantic relations; second, using BabelNet to semantically annotate free text written in any of the covered languages, thus performing multilingual Word Sense Disambiguation in arbitrary languages. This second goal is not addressed in other projects, and represents a step towards the multilingual Semantic Web. Still, the connection to the SW world is weak. In my vision, the next step is to *link data according to ontologies which provide multilingual lexicalizations*, a direction which I believe should be vigorously pursued in the near future and which would see the strong synergy of multilingual Word Sense Disambiguation with Linked Data. Here I do not mean we should create a single, global lexicalized ontology for all purposes. Instead, domain ontologies – even those which are not lexicalized according to a specific language – could be linked to multilingual lexicalized ontologies, which would be used as an interlingua for making Web content accessible to users independently of the language they master. Crucially, the more data will be linked across languages, the better the disambiguation performance (see e.g. [62]). One current problem of the Web of Data, in fact, is the ambiguity (and multilinguality) of labels for Linked Data [27]. W3C standards should be used to encode both the interlingual ontologies and the domain ontologies in a common format, and extensions of existing standards could be developed in order to bring together ontologies and the lexical meanings expressed in a babel of languages.

3.22 Abstract

Sergei Nirenburg (University of Maryland)

1. **What are in your view the most important challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web?** There are many facets to this issue – technological, sociological, bureaucratic, etc. I can comment only on what the field of CL/NLP can contribute to the quality of the multilingual content. From this point of view the main challenge is reaching the level of quality of automatic translation that is acceptable by users. Manual translation is (currently) too slow to support fast turnaround on the Web. The current quality of automatic translation is too low for publication-level texts, though the general opinion is rather more favorable with respect to translation of informal texts, such as blog entries or tweets. The above reckoning is not new and held for the pre-web machine translation applications as well.
2. **Why do current solutions fail short?** See above: the low quality of the translated content. All other shortcomings are easier to overcome but in the final analysis they are not true obstacles.
3. **What insights do we need in order to reach a principled solution? What could a principled solution look like?** In the short term, it is probably best to study the minimum levels of quality that users accept in various types of uses of the multilingual web and try to go after the low-hanging fruit first (not that much else has been going on in NLP over the last 15 years or so). A gradual way of enhancing translation quality is – again, just like in the times of pre-web MT – human-assisted translation. It is not clear that this option would work for the web – after all, response speed is paramount in this environment. But if a successful methodology can be developed and if it can be shown

to keep the costs of human translation down, then human- assisted translation can be a very useful stop-gap solution while more R&D is undertaken to develop high-quality automatic translation. Of course, there is no guarantee that this development period won't take several lifetimes.

4. **How can standardization (e.g. by the W3C) contribute?** I don't think this is a crucial issue either way. In general, I think that standards should evolve and not be propagated top-down by bureaucratic means.

3.23 Under-resourced languages and the MSW

Laurette Pretorius (University of South Africa)

Challenge 1: Under-resourced Languages

In the context of the Multilingual Semantic Web (MSW) the most basic challenges that face Africa, and in particular South Africa, is limited internet access (out of scope here) and the proliferation of mainly under-resourced languages (in terms of financial, linguistic and human resources) with rich cultural diversity and indigenous knowledge systems encoded in these languages. Multilingualism impacts all sectors of African society, viz. public, private, business, education, etc.

As a first step, the Semantic Web (SW) may serve as a safe repository for valuable, already available language data/material by publishing it in the SW. To enable this process of preservation and archiving, the required range of language specific approaches, tools and technologies have to be developed. Care should also be taken to use or adapt, where possible, existing approaches and solutions in order to fast-track these initiatives. Time is of the essence in ensuring that these languages are technologically enabled.

In parallel to these archiving and repository building initiatives, the greater promise and benefits that the SW may offer to Africa, as it moves towards participating in the 21st century knowledge economy, should be immediately pursued. This would require the development of new terminology, state of the art lexical and language processing resources, tools, and technologies for the relevant languages. In addition, a wide and growing range of semantic technologies and ontologies may be required to capture the cultural diversity, the plethora of indigenous knowledge systems and all that goes with moving towards global economic participation and growth.

Challenge 2: Notions for clarification

The notions of "language-independent", "culture-independent", "language-specific", "culture-specific", the conceptual versus the linguistic, and how information about all of these is represented in the MSW, require continued careful consideration – a problem of which the complexity increases with the number and diversity of languages and cultures included.

Challenge 3: Interoperability and ease of use

The representation of the above notions, the information they are employed to encode, and, eventually, the resulting computational semantic artefacts (e.g. localised, mapped, modularised, verbalised ontology, etc.) will have to interoperate and they will have to be

accessible across languages and cultures at a grand scale. At a more modest scale, for the real up-take of emerging semantic technologies and the MSW, it should also be relatively easy for a single user as producer and consumer of specialised content to conceptualise his/her arbitrarily complex interest domains, tasks, and applications, and to use the range of available MSW resources, representation and reasoning tools, etc. to his/her competitive advantage.

Examples of specific functionalities that may be relevant for a wide range of MSW users include:

1. to have access to state of the art support and best practices of knowledge representation;
2. to do sophisticated intelligent searches of specified scope;
3. to delimit the search, access, generation and publication of information in languages of choice;
4. to perform automated reasoning of specified scope and complexity in the MSW;
5. to obtain semantically accurate translations of the retrieved or generated material and of the reasoning results, on request;
6. to provide large-scale automated decision-making support in (multiple) natural language(s);
7. to have access to approaches and tools to evaluate results obtained.

Challenge 4: Continued interplay between natural language processing resources and technologies, semantic technologies and the MSW:

A serious issue in under-resourced languages remains the lack of terminology. The MSW offers unique opportunities in terms of community-based (crowd-sourcing) approaches to, among others, terminology development and moderation, and representations of culture-specific and indigenous knowledge systems. The MSW may serve as an incubator for the continued development of increasingly sophisticated natural language processing and lexical resources for under-resourced languages using new approaches that may emerge due to the availability of rich cross-language support, resources, tools and technologies.

Conclusion:

The balance between appropriate theory and practice will be important in ensuring the sustainability of the MSW. Standards already exist or are in the pipeline for various aspects of the MSW. The ever increasing size and complexity of the MSW will require good standards and best practices towards some notion of integrity of the MSW.

3.24 How to make new domains, new languages, and new information accessible in the Multilingual Semantic Web?

Aarne Ranta (University of Gothenburg)

The crucial idea of the Semantic Web is the formalization of information. The advantage of formalization is that search and reasoning are supported by a well-defined structure: it enables us to work with formulas rather than text strings. This kind of information is easier

for machines to process than free text. But it also involves an abstraction from languages and can thereby serve as a representation that supports multilinguality.

From the perspective of GF[43] [70], a formal structure can be seen as an abstract syntax – as an interlingua, which represents the content in a language-neutral way and makes it possible to produce accurate translations. This perspective is currently exploited in the European MOLTO[44] project (Multilingual On-Line Translation).

The most important challenge for the Multilingual Semantic Web is simply: how to create more of it. The formalization is a wonderful thing when it exists, but it is expensive to create. We the supporters of the idea should try to make it more attractive – both by demonstrating its usefulness and by providing tools that make it easier to formalize web content.

MOLTO's emphasis has been on showing that formalization can give us high-quality automatic translation. MOLTO's show cases have been a small set of domains, ranging from mathematical teaching material to a tourist phrasebook. We have created techniques for converting semantic formalizations to translation systems, which cover up to 25 simultaneous languages. The richest demonstration of the idea is the Multilingual Semantic Wiki, built on top of the ACE Wiki[45] (Attempto Controlled English) by generalizing it to a multilingual Wiki using GF. In this Wiki, every user can edit a document in her own language, and the changes are immediately propagated to all other languages. The translations are thus kept in synchrony at all times. This is enabled by the "master document" which is a formal representation of the content.

The main remaining challenge is: how to make new domains, new languages, and new information accessible in the Multilingual Semantic Web? We need to make it much easier to formalize legacy information, so that we can increase the amount of web content that can be accessed in the rigorous and reliable way. This formalization can benefit from heuristic bootstrapping methods such as machine learning – but it will always involve an element of human judgement, to make sure that the results are correct [46]. The challenge is to find the proper place of human judgement, so that its need can be minimized. The goal is to do more formalization with less effort.

The questionnaire

1. **What are in your view the most important challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web?** Extending the coverage of the Semantic Web and the associated language resources.
2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?** High-quality translation: software localization, e-commerce, technical content, legal information, etc. Also language-based interaction and information retrieval, including mobile speech applications.
3. **Which figures are suited to quantify the magnitude or severity of the problem?** The number of "concepts" involved (1000's in the largest cases of MOLTO; millions in the whole web); the number of languages (up to 25 in MOLTO, thousands in the world).
4. **Why do current solutions fail short?** Too much and too boring human work is needed; its usefulness has not been convincingly demonstrated.

[43] GF, Grammatical Framework, http://www.grammaticalframework.org/
[44] MOLTO, Multilingual On-Line Translation, http://www.molto-project.eu/
[45] ACE Wiki, http://attempto.ifi.uzh.ch/acewiki/

5. **What insights do we need in order to reach a principled solution? What could a principled solution look like?** We need to understand what is easy (automatic) and what is di cult (needs human labour). We need to create logical and linguistic resources of high quality, coverage, and reusability, with completely free access.
6. **How can standardization (e.g. by the W3C) contribute?** By eliminating duplicated work. For instance, if there is an ontology and multilingual lexicon of fish names, everyone can use it off the shelf and don't need to build their own from scratch. Having a common format is less important. It is OK to have different formats as long as they are fully specified and can be converted to each other.

3.25 Abstract

Kimmo Rossi (European Commission, DG CONNECT)

This seminar comes at an interesting point in time, when we at DG CONNECT are defining the European Data value chain strategy, in view of establishing orientations for the first work programmes of Horizon 2020 (H2020), which is the next framework funding programme supporting research and innovation in ICT. Very soon after the seminar, we will launch consultations of stakeholders (researchers, industry, civil society, administrations) that will feed into the process. In early 2013 we need to have first stable topical orientations for the first phase of Horizon 2020. I expect this seminar to detect, define and refine some of the key methodological and scientific issues and challenges related to the data challenge in general, and the linked data/semantic web/text analytics challenge in particular. If it does, it will provide valuable input to the process of defining H2020 and other related programmes. With the recent reorganisation by which DG INFSO was rebaptized DG CONNECT, the units responsible for information management (E2) and language technologies (E1) were combined into one unit. This created a single pool of over 100 ongoing research and innovation projects with over 300 MEUR funding, mobilising more than 1500 full time equivalents of Europe's best brains to break the hard problems that currently hamper effective use and re-use of online content, media and data. We try to use this pool of ongoing R&I projects also as a tool to bridge into the future research agenda in H2020, still taking shape. So, there are plenty of resources, the challenge is to make them converge and contribute to a common effort.

I also hope this seminar to be an opportunity of deepening the views and ideas that were presented at the Dublin workshop of the MultilingualWeb project, especially the 1st day, dedicated to the theme "Linked open data"[46].

Below are my personal views concerning the specific questions:

1. **What are in your view the most important challenges/ barriers/ problems and pressing needs with respect to the multilingual access to the Semantic Web?** In general, the concept of Semantic Web (accessing and addressing the web as a database) requires precise, fast and robust automated text analysis capabilities which are not there, especially for languages other than English. Since the increasing majority (75%

[46] http://www.multilingualweb.eu/en/documents/dublin-workshop/dublin-program

or so) of Web content is in languages other than English, the text analysis bottleneck gets worse over time.

2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?** Huge potential benefits are currently missed because of lack of semantic tagging and linking of documents. Such benefits could be reaped in various sectors, but the most obvious are publication, communication, marketing, intelligence, tourism, biomedical, administration – any industry or activity which relies on fast access to relevant facts from large numbers of textual (human language) sources, some of which are static documents, other streams of data.
3. **Why do current solutions fail short?** I may be wrong, but I have a feeling that state of the art information extraction has not been systematically utilized in efforts to link data, for example in efforts like DBpedia. Also, there is too much hand-crafting, domain-adaptation, and other tweaking that is not replicable and robust. When the characteristics of the underlying data change, such schemes risk becoming obsolete. Another thing is that the performance of automated tagging, information extraction, named entity recognition etc. are heavily dependent on the language (and completely missing for some languages). Finally, methods and solutions should be able to cope with
4. **What insights do we need in order to reach a principled solution? What could a principled solution look like?** In general it seems like a good idea to improve the generality, robustness, adaptability, self-learning/self-training nature of methodologies (because of the changing nature of the data and tasks).
5. **How can standardization (e.g. by the W3C) contribute?** Rather than prescribing solutions, standardisation should provide a flexible framework and practical tools to cope with the diversity of data types, formats, languages, business practices, personal tastes etc. In general, the Web seems to defy any major standardisation or regulation attempts. The best practices will establish themselves organically (i.e. superior solutions will automatically prevail), and this should also be the guiding principle for standardisation. Demonstrating, documenting and sharing best practices is an effective way of setting standards (like W3C does). What realistically can be standardised, should be standardised. For the rest, systems and solutions should be designed to cope with multi-standard (or standardless) reality.

3.26 Sustainable, organisational Support for bridging Industry and the Multilingual Semantic Web

Felix Sasaki (DFKI/ W3C-Fellow)

1. **Challenges** A huge challenge for the multilingual Semantic Web is its separation to other types of multilingual content. In industries like localization, terminology or many areas of linguistic research, creation of multilingual resources is resulting in fast amounts of content in many languages. Unfortunately, this content cannot be easily re-used on the multilingual Semantic Web, and resources created within the multilingual Semantic Web are rarely part of industry applications. The underlying issue is partially one

of information integration; this problem is already tackled via formats like NIF[47] and LEMON[48], which help to re-use and combine tools and resources (e.g. lexica) for the creation of multilingual resources. However, another part of the issue is the topic of content creation and localization workflows, which is different in industries compared to the multilingual Semantic Web. This difference can be characterized as follows:

- In localization and content creation, multilingual resources are being created in complex workflows with many organizations involved. In the multilingual Semantic Web, multilingual linked open data is rather created on a project specific basis by research groups. This leads to uncertainty with regards to the quality and maintenance of the data.
- Trustworthiness and quality of data is an important aspect of workflows in localization and industrial content creation: e.g. the localization of medical information needs to take national and — esp. in translation scenarios -– international regulations and quality measures into account. Data currently available on the multilingual Semantic Web not only differs highly in terms of quality; an evaluation of the quality itself (level of maintenance,trust of content creators etc.) is hard to achieve.
- Closely related to trurstworthiness are legal aspects of linked data, e.g. what data can be re-used with what kind of license. Without such information, data from the linked open data cloud will not be re-used in large scale industrial scenarios.

2. **The role of the problems in industry practice** The above problems matter in practice since so far the usage of linked open data in areas which are inherently multilingual, that is content creation and localization, is rather low. This does not only have to do with current technical solutions for the multilingual Semantic Web itself: as Jose Emilio Labra Gayo (2012)[49] demonstrates, in the current technical infrastructure there are already means to create multilingual information within linked open data; unfortunately these are rarely used, and the actual amount of multilingual data in the Semantic Web is rather low.
3. **Why do current solutions fail?** Technical advances are a mandatory part of a solution to the problem, see e.g. LEMON and NIF mentioned above. Nevertheless, failures are also due to organizational issues. An example of this situation are language identifiers and what I will call the "zh" problem. "zh" is the language identifier for Chinese created as part of ISO639-1. The first edition of ISO639-1 was approved 1967; here "zh" is described as an identifier for Chinese in general. However, for decades, it has mainly been used for identifying content in the Mandarin language. With the creation of ISO639-3, Mandarin received its own language identifier "cmn". "zh" was defined as a so-called macrolanguage, acknowledging that there is no single Chinese language. "zh" now is a macrolanguage covering closely related languages like Mandarin or Hakka.
   The new role of "zh" leads to the situation that a language tag in existing content like "zh-tw" can have several interpretations: the macrolanguage Chinese spoken in Taiwan, Chinese in the traditional script, or Mandarin in Taiwan.
4. **What insights do we need in order to reach a principled solution?** The lesson to be learned from "zh" is that what is needed are not only multilingual resources, e.g.

[47] NIF (NLP Interchange Format) http://nlp2rdf.org/
[48] LEMON (LExicon Model for ONtologies), being standardized in the W3C Ontology- Lexicon community http://www.w3.org/community/ontolex/
[49] Best Practices for Multilingual Linked Open Data. Presentation at the 2012 workshop "The Multilingual Web — Linked Open Data and MLWQLT Requirements", Dublin. See http://www.multilingualweb.eu/documents/dublin-workshop/dublin-program

the identifier "zh", or advances in technical solutions. In addition, organisational and workflow information about the context of content creation and applications need to be established.
Various industries (libraries, terminologists, general software companies, Web centered companies and multimedia companies) are working closely together to solve problems which arise from the above situation, that is: to make sure that in a given workflow, "zh" can be interpreted in the appropriate manner.
Currently the community of multilingual Semantic Web is not part of the related organizational structures, which creates barriers between the multilingual Semantic Web and other, inherently multilingual industries. The bad news is that there is no general, principled solution to resolve this. But we can make steps and long-term plans which will help to address the problem.

5. **How can standardization contribute?** Standardization is one important part to solve the workflow problems described in this abstract. The community building needed to solve the "zh" problem for the industries mentioned above mainly is happening in standardization bodies like the IETF, the Unicode consortium or W3C. The aformentioned efforts of lemon and NIF show that the NLP community is making efforts into the direction of standardization. The W3C MultilingualWeb-LT Working Group[50] is another effort with special focus on community building involving industries, making sure that there is awareness for issues mentioned under 4).
Nevertheless, the conclusion is that this is not enough: what is needed is an approach also towards research in which integration with industry workflows is not an aftermath or part of separate projects. Sustainable, institutional support for bridging the workflow related gaps mentioned in this abstract is needed. We should put an effort in describing research and (product) development not as two separate lines of action, but as closely integrated efforts. How this sustainable institutionalization of multilingual research and innovation should be framed in detail and how it should be worded in upcoming research programs like Horizon 2020, is an important and urgent topic. The Dagstuhl seminar should help to move this discussion forward, also just by bringing the relevant players together.

3.27 Abstract

Gabriele Sauberer (TermNet)

1. **Most important challenges/barriers/problems and pressing needs with respect to the multilingual access to the Semantic Web:** The main problem of the Semantic Web and the Multilingual Semantic Web (MSW) alike is the imbalance of its players and drivers, i.e. the lack of diversity: they are mainly male, white, academic, IT-focussed and aged between 20 and 45.
It's not only a matter of dominance of "English language and Western culture" as correctly stated in the Synopsis of the Dagstuhl Seminar, it is much more a matter of a global digital divide: a divide between men and women, between age groups, social

[50] See http://www.w3.org/International/multilingualweb/lt/ for further information.

classes, between disciplines, subject fields, traditions, cultures, information and knowledge rich and information and knowledge poor, between literates and illiterates, experts and non-experts, etc.
Thus, one of the pressing needs with respect to the MSW is to address the lack of diversity and to overcome the global digital divide.

2. **Why does the problem matter in practice? Which industry sectors or domains are concerned with the problem?** Lack of diversity in working together to build a Semantic Web that matter for all citizens is, to my mind, one of the main reasons why MSW got stuck – technically, economically and socially.
The word-wide acceptance of the Semantic Web is a key issue of its development and survival. People all over the world understood the benefit and practical advantages of mobile phones in their lives very fast.
What's in it for all of us when using semantic web technologies and smart phones is the core message to be brought home by the drivers of the MSW.
There are many industries which can help in overcoming language and national barriers: the language industries, education and training industries, Information and Communication Industries, etc. "Facilitating semantic access to information originally produced for a different culture and language" as mentioned in the synopsis is to be avoided, to my mind, not fostered. Why? Because goal and vision of MSW should be to empower people to contribute to the MSW by their own in their own languages, not being restricted to adept and localize foreign content.
Terminological methods, tools, trainings and consultancy services are, to my mind, key technologies and basic knowledge prerequisites to contribute to problem solutions.
3. **Which figures are suited to quantify the magnitude or severity of the problem?** The industrial relevance of MSW and its barriers is high-lighted at page 3 of the synopsis, e.g.:
Especially in knowledge-intensive domains, such as finance, biotechnology, government and administration etc., the ability to interface with Semantic Web or Linked Data based knowledge repositories in multiple languages will become of increasing importance. In finance, knowledge repositories will be build up of company-related information, i.e. in terms of finance, markets, products, staff, all of which will be curated and accessed in multiple languages.
No doubt, we are talking here about a hundreds of Billions Dollar, Euro, RMB etc. business/ losses.
4. **Why do current solutions fail short?** *Because the current solutions are no sustainable, future-oriented and no global solutions: They lack of creativity and innovation, caused by lack of diversity. It's just more of the same, provided by the same players (see question 1).*
5. *What insights do we need in order to reach a principled solution? What could a principled solution look like?* What we need is, to my mind, exchange of insights at all levels, representing the diversity of current and future users of MSW (all genders, all age and social groups, all regions, cultures, disciplines, literates, illiterates, etc.).
To seize opportunities for new insights is crucial and simple, but not easy: We mainly need to overcome the barriers and restrictions in our minds, in our ethno-centric attitudes and behaviour.
A principled solution could be to make diversity a main principle of the Semantic Web and the MSW: With this new and lived principle, diverse teams and diverse expert and working groups, guided by communication and terminology experts could make the vision

of a real multilingual and multicultural Semantic Web come true.

6. *How can standardization (e.g. by the W3C) contribute?* Standardization organizations and their Technical Committees as well as the W3C can contribute to the principled solution by developing, issuing and promoting respective standards and guidelines to organize and support diversity as main principle of the Semantic Web and the MSW.

3.28 The Translingual Web – A Challenge for Language and Knowledge Technologies

Hans Uszkoreit, Saarland University

The web becomes more multilingual every day and so does the society of web users. This is not surprising since many large organisations and societies are already multilingual in their processes and constituencies. Multinational corporations, international organisations, professional associations as well as national and regional societies such as the European Union, South Africa, India and Russia often work in many languages.

The World Wide Web has become the predominant medium for sharing information, knowledge and artistic content. The Web is also turning into the universal platform for an endless number of services that extend the static content of the web by functionalities using or modifying this content or just utilizing the Web protocols for all kinds of transactions.

There is a strong demand for making the fast-growing multilingual web also truly crosslingual so that the global medium, which unites all the contents and services in more than thousand languages, would also eventually become the instrument for overcoming all language barriers. So-called globalized websites and web services today offer contents and services in 20 to 35 languages. Such websites are hard to maintain, especially if the contents grow and if the services depend on reliable translation. Google translate offers translations from and into 57 languages. The popular translation facility is an invaluable service making Internet content accessible to large parts of the world's population that would otherwise be deprived of the Web's blessings. However, well-known quality deficiencies of today's translation technology limit the role of the existing large online translation services as the universal crosslingual information and communication hub.

The semantic web is an ambitious program driven by a powerful vision and a promising approach toward a web of knowledge-based services that become much less dependent on human language and therefore also on human languages. If the entire Web could be encoded in semantic representations that are language-independent and suited for automatic reasoning, the crosslingual function of the Web would boil down to multilingual access facilities. The main challenge for such a natural language interface would be the understanding of spoken or written queries and commands. Compared to this unsolved central problem of language processing, the non-trivial task of responding in any requested language is comparatively simple, as long as the semantic web services select the appropriate output in a structured representation.

However, in the foreseeable future we will not witness a Web in which all content is represented in a disambiguated structured semantic representation. At best, a growing layer of semantic web content and services will sit above the wealth of unstructured content, containing large numbers of links into the texts (and possibly also other media) that let the

extracted knowledge also serve as metadata for the unstructured information base. Two observations: (i) the evolution of this semantic layer proceeds in ways not quite foreseen by the early visionaries of the Semantic Web and (ii) the resulting large heterogeneous distributed metadata repository may soon become the most important research and technology resource for getting at the meaning of so-called unstructured data, especially texts for which such metadata do not yet exist.

After more than 50 years of research with human language processing, the scientific community has learned from a combination of experience and insight that in this discipline there are no miracles and no free lunches. Neither teraword data nor a century of grammar writing, neither fully automatic learning nor intellectual discovery and engineering alone will suffice to get us the technology that reliably transforms language into meaning, or one language into another language. It will not even give us the tools for correctly and exhaustively transforming every meaningful linguistic utterance into its paraphrases in the same natural language. Even treebanks including parallel treebanks for several languages with their sophisticated structural information will not provide a sufficient empirical basis for the maturation of language technology. The recognition of the need for semantically annotated textual data keeps growing. Even large semantic resources such as Yago or Freebase whose knowledge units are not directly linked to the texts they came from, have proven highly valuable in language technology research, especially in information extraction.

But the need for semantic resources also includes translation technology. After hierarchical phrase-based and syntax-based translation, semantics-based statistical translation will become the next big trend in MT. On the other hand, knowledge technologies will not be able to get into full blossom either without the evolutionary upgrade path from the textual knowledge representation to the semantic metadata layer.

It seems that the prospects of language technologies and knowledge technologies are inseparably tied to together. Since each of the two technology disciplines needs the other one for reaching fruition, only a complex mutual bootstrapping process around a shared stock of data, tools and tasks can provide the base for effective evolution. A Multilingual Semantic Web layer could become the shared resource of data and tasks for this process, if the Multilingual Semantic Web indeed becomes a core component of the envisaged Translingual Web, it could also incorporate the services needed on the NLP side. Such services would not only cater to research but also gradually fill the growing cross-lingual needs of the global multilingual society.

Both in language technologies and in knowledge technologies research has become much more interconnected and collective in nature. As in the natural sciences and other engineering disciplines, new forms of collaboration and sharing have developed. The sketched bootstrapping will require additional efforts in sharing challenges and resources. How could such efforts be triggered and organized? For European language technology community, some important planning steps toward large-scale cooperation have been taken.

Coordinated by the Multilingual Europe Technology Alliance (META), the European language technology community together with technology users and other stakeholders has drafted a Strategic Research Agenda (SRA) [13], in which the special needs and opportunities for language technology in our multilingual European society are outlined that should drive our research. From these findings, the SRA derives priorities for large-scale research and development as well as a plan for implementing the required massive collaboration. Among the priorities are three solution-driven research themes that share many technologies and resources. All three priority themes are tightly connected with the topic of the Dagstuhl Seminar: (i) a cloud computing centered thrust toward pervasive translation and interpretation including

content and service access in any language, (ii) language technology for social intelligence supporting participatory massively collective decision processes across social, linguistic and geographic boundaries and (iii) socially aware proactive and interactive virtual characters that assist, learn, adapt and teach.

As a means for experimentation, show-casing, data-collection, service integration and actual service provision, a sky-computing based platform for language and knowledge services is planned that needs to be realized through a cooperation between industry, research and public administration. This platform would be the natural target for experimental multilingual semantic web services.

The interoperability of the services will to a large degree depend on the success of ongoing standardization efforts as conducted in collaborations among research, industry, W3C and other stakeholders in the framework of the initiatives Multilingual Web and its successor LT-Web.

Besides the valuable exchange of recent research approaches and results, the Dagstuhl seminar could play an important part in a better linking of the following five research areas in planning:

1. Semantic Web Standards and methods
2. Linked open data and other knowledge repositories
3. Multilingual Web Standards
4. Translingual (Web-based) Services
5. Information/Knowledge Extraction

In addition to new developments from the META-NET/META community (visions, strategic research agenda, schemes for distributed resource sharing) I will try to contribute experience and perspectives to this endeavour from two specialized research strands: (i) minimally-supervised and distantly supervised n-ary relation extraction and (ii) quality-centered translation technology.

3.29 Problems and Challenges Related to the Multilingual Access of Information in the Context of the (Semantic) Web

Josef van Genabith (Dublin City University)

The topic of the Dagstuhl Seminar is broad, especially as the "Semantic" part in the title is in brackets, which could suggest optionality, as in "Web" or "Semantic Web". Accordingly, the notes below quite broad (and rambling).

1. **Challenges** Challenges include (and go well beyond) a mixed bag of related philosophical (knowledge representation, epistemological, reasoning), granularity, coverage, multilinguality and interoperability challenges.
   Philosophical the semantic web aims at capturing knowledge, mostly in terms of concepts and relations between concepts, to support automatic access to and reasoning over knowledge. However, the base categories are not clear. What is a concept? Do concepts exist independently of culture, language, time? Are concepts conventionalised, political, or even ideological constructs? Is it a matter of degree, is there a spectrum with extreme ends with pure concepts on the one hand and completely culturalised concepts on the

other. If so, how do we know what is where on the spectrum and how does this impact on computation? Are ontologies in fact folksonomies etc.? Do we need to bother? Yes, as multi-linguality (amongst others) shows that concepts are not as universal as perhaps assumed. Reasoning (beyond relatively simple applications) is extremely challenging both computationally and conceptually: reasoning with events, temporal information, hypothetical information, contradicting information, factivity, sentiment, probabilistic information, causality, etc. Maybe ontological information (Semantic Web) should be extremely confined/demarcated (factual, as accepted by a culture) backbone component feeding into to a more general inferencing process (link with NLP).
Granularity: for many applications a shallow ontology is quite useful. There is a question when to use/compile shallow or deep/detailed ontological information and for which purpose (fit-for-purpose)? What is not covered by ontological information?
Evaluation: how do we evaluate ontological information? Is there an abstract measure, or is it just in terms of some usefulness criterion given a task (extrinsic evaluation)?
Coverage, quality, noise: content (un-, semi and structured) is exploding on the Web with ever increasing volume, velocity and variety. How do we obtain ontological information: manually compiled, automatically compiled from (semi-) structured input (tables etc.), or from raw text (through NLP/IE)? We need to negotiate the engineering triangle: cheap, fast, quality (you can only have two out of the three at any one given time).
Multilinguality: most of the Semantic Web is in English. Multilinguality raises issues including culture specificity, mapping between ontological information resources (which is quite alien given the often un-articulated background assumption that ontologies are about concepts that may help transcend languages and cultures), overall structure for ontological information (one concept for each culture and "translations" between them, a single one with sub-categorizations: Chinese, Arabic, Western ....)?
Interoperability: multi-lingual, -cultural, -granular, -redundant, -domain, -... how do we make this all play in concert? How do we make NLP/IE and Semantic Web interoperable? They should be able to contribute much to each other, in fact (some of) the trouble starts when you make each one of them do it all on its own ....

2. **Why does the problem matter in practice?** We need to capture knowledge to support technology mediated access to and interaction with knowledge.
3. **Figures that quantify the problem:** Content velocity, volume and variability is steadily increasing: rise in User Generated Content (UGC) with Web 2.0 move of users from passive consumers to active producers of content. English is rapidly loosing its role/status as the language of the web. Most growth in the web is from emerging economies.
4. **Why do current solutions fail short?** The trouble is there are different kinds of knowledge (of which the Semantic Web captures some), the volume of knowledge is constantly increasing (of which the Semantic Web captures some), knowledge is dynamic (i.e. constantly changing, updating) (of which the Semantic Web captures some), knowledge is encoded in different formats (un-, semi- and structured) (of which the Semantic Web captures some) and different languages (of which the Semantic Web captures some).
5. **Principled solution:** In my view making Semantic Web and NLP play in concert supporting each other is one of the greatest challenges. NLP can provide scalability, the capacity to address content/information velocity (frequent updates), volume and variety. Semantic Web can provide knowledge guiding NLP. NLP can help bridge language barriers.

6. **Standardisation:** Full standardization is difficult to achieve. It may be more realistic to aim for some kind of interoperability of heterogeneous sources of information/knowledge.

3.30 Towards Conceptually Scoped LT

Jeroen van Grondelle (Be Informed)

This contribution aims to provide an industry perspective on the multilingual semantic web and tries to answer three questions: What is the semantic web used for today, why is (natural) language important in the context of the semantic web and, only then, how could that guide the development of the multilingual semantic web.

A Semantic Web of Unconsolidated Business Constraints

For reasons that probably differ from the original semantic web vision, companies and governments alike are embracing semantic technology to capture the information they use in declarative, interoperable models and ontologies.

They move beyond the data and capture the policies and definitions that govern their daily operations. By choosing the right conceptualizations for business aspects such as products, business processes and registrations, they manage to use these ontologies to design their business, agree on the terms used to communicate its intentions and execute the required applicative services that support it.

The semantic web stack of ideas and technologies fits them well. Although the problem does not have global web scale, they benefit from the unconsolidated nature of the semantic web technologies when capturing aspects across different organizations and departments. The vocabularies that emerge when modeling this way have proven very valuable in communicating policy between all stakeholders.

A Lingual Semantic Web for Humans AND Machines

Although a lot of emphasis lies with machine's ability to interpret and reason with ontologies on the semantic web, we believe that human's role is crucial.

In our experience, ontologies can be useful to organizations throughout the policy lifecycle: From shaping the organization by drafting policy candidates and choosing the policy that is expected to meet the goals best, implement policy in business processes and applications, execute policy and evaluate policy by reporting and collecting feedback. When ontologies are used to facilitate these processes, users will interact with the ontologies in many ways: Domain experts and business users need be the owners of the ontologies and therefore need to create, review and validate the ontologies. and users need to interact with and understand the services based on the ontologies, ranging from online forms to complex process applications.

Language plays an import role in supporting the different forms of dialog needed. Often the boundaries of dialog supported by a specific language technology are specified at a lingual level, bound by lexicons and types of lexical constructs understood by the technologies etc. When used in the context of the semantic web, we believe that the boundaries of dialog need to be specified in two dimensions: The domain that is discussed in dialog, typically represented by an ontology, and the task context of the dialog.

Often the implicit task model under semantic web applications is limited to querying and presenting instances in the ontology, or aggregates of them. Both the original semantic web

vision and the businesses using semantic technology today require more complex tasks to be performed based on ontologies, and more diverse dialogs as a consequence.

For example, an ontology might be use to capture policies on eligibility for a permit [IND]. Typical task models throughout the policy lifecycle might require dialogs that

- Speak of the domain in general or that speak about a specific permit application;
- Allow citizens to apply for a permit, providing all required facts and receiving feedback on the consequences of the information they provide;
  - Discuss an application with an official or with the applicant himself;
- Support experts to validate the ontology and maintain consistency;
  - Represent not only the ontology, but also represent generated contradictions to trigger feedback.
- Support what if analyses, that describe possible changes in the legislation or in the population of applicants and the consequences on the permits issued.

All these types of dialog require ingredients both from the domain ontology and from the task model, probably at both a semantical level and the lexical representation level. Challenges in LT might include representing the different aspects of such a task model, and how to decouple it in an orthogonal way from the domain ontology.

Use Cases and Challenges for Multilingualism

Given the importance of language in interacting with the semantic web, it is clear that multilingualism is crucial when applying semantic web technologies at serious scale and in international context. Apart from providing transparent access to services and dialogs based on ontologies, multilingual capabilities of the semantic web are important for sharing and reusing ontologies and facilitate collaboration across languages in the process of creating and agreeing on ontologies that capture international standards.

The task orientation introduces requirements to multilingualism beyond translation of all concepts in both dimensions. There are a lot of lingual, even cultural aspects to having a dialog, such as when to use formal forms, what are the preferred ways to ask, confirm and give advice for instance.

Conclusion

Most language is spoken in dialog, in the context of a task or a common goal and in a certain domain. Language technology that incorporates the conceptualizations of all these aspects and is able to generalize across languages has useful applications today in lots of areas, including the business processes of both companies and governments. That may be a first step to making the semantic web vision a reality: a web of intelligent agents and services based on unconsolidated, distributed ontologies, created and owned by domain experts.

3.31 What is the current state of the Multilingual Web of Data?

Asunción Gómez Pérez & Daniel Vila Suero (Universidad Politécnica de Madrid)

1. **Motivation**

   The Semantic Web is growing at a fast pace, recently boosted by the creation of the Linked Data initiative and principles. Methods, standards, techniques and the state of

technology are becoming more mature and therefore are easing the task of publication and consumption of semantic information on the Web.

As identified in [34] this growing Semantic Web offers an excellent opportunity to build a multilingual "data network" where users can access to information regardless the natural language they speak or the natural language the information was originally published in. But it also creates new research challenges and presents some risks, being the most significant one the creation of, what the authors describe as, "monolingual islands" – where different monolingual datasets are disconnected to datasets in other languages.

Having this in mind, we pose ourselves two simple questions:

- Are we able to devise representative statistics and findings that could help us to shed some light on the current state of the Web of Data with respect to the use of natural languages?
- Can such statistics and findings serve us in the development and testing of new tools, techniques and services that could help to overcome the aforementioned challenges?

The preliminary work we present here represents an effort to gather useful information and resources that could help to face already identified research challenges, discover new ones and provide a base for discussion within the research in Multilingual Semantic Web.

Our initial objective will be to answer questions like:

- What is the distribution of natural languages on the Web of Linked Data?
- To which extent are language tags used to indicate the language of property values?
- Which domains are predominantly mono/multilingual?
- What is the distribution of cross-lingual links vs. monolingual links?
- How are cross-lingual links established (e.g. owl:sameAs)?
- Are we able to identify monolingual datasets not connected to data in other languages and thus conforming "monolingual islands"?
- Do mono/multilingual datasets organize themselves into clusters with respect to the used natural languages?

The remainder of the text gives an overview of the resources we count on in order to set up an environment for our study (Section 2), and the methodology of our study (Section 3).

2. **Resources**

With the increasing amount and heterogeneity of data being published to the so-called Web of Data, in the recent years we find (1) several measurements and empirical studies of web data, (2) crawled datasets providing representative subsets of the web of data.

Regarding (1), we find several works such as [22, 17, 27]. However, most of them do not take into account the multilinguality at all (e.g. [22, 17]) or do it to a limited extent (e.g. [27]). In [27], Ell et al. introduce a set of label-related metrics and report their findings from measuring a subset of the Web of Data[51] using the proposed metrics. One of these metrics is the multilinguality of the labels. More recently we find the LODStats[52] initiative that aims at gathering comprehensive statistics about datasets adhering to the RDF found at thedatahub.org. In the website we find statistics about languages[53], however it remains unclear how these data are gathered and it lacks absolute numbers

[51] The authors used the Billion Triple Challenge Dataset 2010 (see http://km.aifb.kit.edu/projects/btc-2010/).
[52] http://stats.lod2.eu
[53] http://stats.lod2.eu/languages

that could help to analyse for example the distribution of usage of language tags (i.e. to which extent are language tags used to indicate the language of the property values).
Regarding (2), since 2009, a number of crawled corpora are being made publicly available in order to facilitate the analysis and characterization of web data. One of the most significant examples is the Billion Triples Dataset, already used for a number of studies (e.g. [27, 23]). More recently, we find the "Dynamic Linked Data Observatory"[54], a framework to monitor Linked Data over an extended period of time [45]. In [45] the authors discuss the strengths and weaknesses of two perspectives of the web of data (the BTC dataset and of what they call the CKAN/LOD cloud metadata[55]) and propose high-quality collection of Linked Data snapshots for the community to gain a better insight into the underlying principles of dynamicity on the Web of Data.
Given that we find very few work on analysing the Web of Data from the perspective of multilinguality, in the next section we propose a methodology for performing our study.

3. **Method and rationale**
After analysing aforementioned corpora and performing several analysis, our initial candidate for extracting statistics and issuing questions will be the Dynamic Linked Data Observatory, being the most important reasons behind our decision the following: (1) it has reasonable size, (2) it is updated frequently updated so we can periodically run our analysis, (3) it tackles some of the issues found in BTC and LOD.
After selecting the corpora, we have set up an infrastructure based on Apache Hadoop and Apache Pig that allows us to periodically analyse the data (i.e. every time a new corpora gets published) and run the different questions that we want to answer.
Having the dataset and the infrastructure, the method will be the following: (1) every time we want gather statistics on a new feature, we create a simple script and store it, (2) every time a new "observatory corpus" is published the stored scripts are executed, (3) the results can be analysed and published for the community.
We are currently in the first steps of this effort, but we are able to share some of our results and we would like to make this a community effort where researchers can suggest new studies and perspectives. During the seminar we would like to share some of our results, validate our current questions and gather new ones from other interested participants.

4. **Acknowledgements**
This work is supported by the Spanish Project TIN2010-17550 for the BabeLData project.

3.32 Exploiting Parallel Corpora for the Semantic Web

Martin Volk (University of Zurich)

The biggest challenge in multilingual access to the web is still the limited quality of machine translation. This may sound like a somewhat trivial observation, but it clearly points to the core of the problem. Machine translation has made big progress. Because of statistical machine translation we can build translation systems quickly for many language pairs when

[54] http://swse.deri.org/DyLDO/
[55] The CKAN(Comprehensive Knowledge Archive Network) repository contains a group lodcloud which is the one used in the creation of the LOD cloud.

large amounts of translated texts are given for the languages and domains in question. The quality of machine translation in many application areas is good enough for profitable post-editing rather than translating from scratch. But the quality is often still a problem when using the machine output in other applications like cross-language information extraction.

Large collections of translated texts (parallel corpora) are the crucial prerequisite for advancing not only the field of machine translation but also any approach to automatically learn cross-language term correspondences and to verify and disambiguate ontological relations for each language. After all large text collections are the sole basis for automatically extracting semantic relations on a large scale.

Therefore I see it as our most important task to collect parallel corpora, to encourage more people to provide parallel corpora and to support any initiative for the free access to a wide variety of parallel corpora.

Still, we shall not forget that statistical approaches to translation and to ontology building are not an option for many lesser-resourced languages (which account for the vast majority of the languages spoken on the planet today). In order to work against the widening of the technological gap, we need to find viable approaches to build bilingual dictionaries with large and lesser-resourced languages, to collect special corpora for such language pairs and to include these lesser-resourced languages in our ontology building efforts.

All activities towards multilingual information access, in the web in general and in the semantic web in particular, will benefit many types of industries and organizations: the media industry (newspapers, TV stations, film industry), the manufacturing industry (user interfaces, user manuals, documentation), trade and commerce (communication, agreements, contracts), administration (law texts, court decisions, patents), tourism, science and education.

4 Working Groups

In this section we summarize the discussion topics and main outcomes of each of the group working sessions as well as of the final session on Friday morning. Figure 3 shows some snapshots of groups during their discussion.

Figure 3 Working together during the group sessions.

The group working sessions were the following ones:

4.1 Bootstrapping a Translingual Web (Day 1)

We agreed the vision of a translingual Web to consist of a web that mediates between different languages, and which supports the location and aggregation of facts independently of the language in which they are expressed, allowing users to pose a query in any language and receiving answers translated into her language from other languages. The translingual Web relies on language-independent representation of all the content available on the Web, with the possibility to map language-specific queries into these language-independent representations and surface-realizing these results in the language of the user.

The creation of a translingual Web constitutes a chicken-and-egg problem. On the one hand, if robust, deep and accurate semantic analysis for different languages would be available, then language-independent representations of content could be generated automatically and the translingual Web would be essentially in place. On the other hand, if all the content of the Web would be formalized in Semantic Web languages, then the translingual Web would be already there, as all the content would be captured with respect to language-independent Semantic Web vocabularies. Cross-lingual query processing would essentially consist of mapping queries in any language into semantic Web vocabularies and translating back the answers into the language of the user.

As both solutions are not there yet, the only reasonable approach is to assume that a translingual Web can be bootstrapped by incrementally i) improving the available technology for semantic analysis of natural language in multiple languages and ii) increasing the amount of Semantic Web content available.

In bootstrapping a translingual Web, the following communities need to be involved:

- **Ontology engineering and reasoning:** The ontology engineering and reasoning communities could develop reasoning and inferencing techniques allowing the detection of inconsistencies across facts expressed in different languages at large scale.
- **Linked Data:** The linked data community could contribute to the endeavor by enriching linked datasets with multilingual features and giving more prominence to linguistic information.
- **Ontology Alignment:** could contribute by developing techniques that support the alignment of ontologies, vocabularies and datasets across languages.
- **Ontology Learning:** the ontology learning community could contribute by developing approaches that can learn language-specific ontologies and detect gaps and modelling differences across different taxonomies.
- **Question Answering:** could contribute by developing methods that can answer questions in different languages on the basis of the growing amount of Linked Data, exploiting cross-lingual mappings and correspondences.
- **Machine Translation:** a current hot topic in MT is how to exploit semantic information to improve machine translation. It needs to be explored whether the type of semantics contained in Semantic Web ontologies can be exploited for MT purposes.
- **Information Extraction:** Extracting information at web scale for all the various semantic web vocabularies is a challenging task for current information extraction algorithms. Developing algorithms that can reliably extract relations and entities for a high number of Semantic Web ontologies and vocabularies is an important challenge to focus on in the next years.
- **Textual Entailment:** The textual entailment community could contribute by focusing on developing entailment approaches that work across languages and can also discover inconsistencies.

A good starting point for research on the translingual Web would be to transform Wikipedia into a translingual resource. The benefit of Wikipedia is that it is a free, multilingual resource that has textual content in multiple languages as well as structured content (in DBpedia). Further, it is highly interlinked to other datasets. Users should have an immediate benefit of such an endeavour. One could think of two application scenarios: question answering and factual error detection in Wikipedia. In the first scenario, one would allow users to formulate a query in their own language. Examples of queries could be looking for information on a foreign soccer club, perform research on the legislation of a foreign country or compare information (e.g. the energy mix used) in different countries.

In the factual error detection use case, one could compare the facts extracted from textual Wikipedia data in different languages to the data available in infoboxes in different languages. Additionally, one could try to find additional facts that do not fit into the current infobox templates.

4.2 Content Representation and Implementation (Day 1)

The discussion in this group was structured around three topics:

- What can SW offer to NLP, what NLP to SW?
- What evidence is there that this is happening?
- What are the roadblocks?

One of the promises is that Semantic Web can improve natural language processing. Indeed, semantics can improve NER and contribute to reduce sparsity (possibly in MT and parsing?). The Linked Open Data Cloud can support complex Q&A. However, one observation is that the Semantic Web seems to be almost not acknowledged by the NLP community. The awareness is almost non-existing.

On the other hand, NLP can definitely support the Semantic Web, allowing it to scale in terms of content by providing information extraction, annotation, algorithms for term extraction as well as taxonomy and ontology construction, etc. There is ample evidence that this is happening, as the European Semantic Web Conference has an NLP track (since 2011).

Overall, we need more of this synergetic work happening. However, while triples seems to be a reasonable representation from a data-oriented point of view, the expressivity of current triple-based SW languages such as RDF(S) and OWL are not sufficient from an NLP perspective, being silent about how to represent tense, aspect, plurality, modality, quantification, propositional attitudes, etc. However, we need to be pragmatic here and live with the representational formalisms we have in the SW. SW needs to clearly move away from symbolic and crisp-KR and move towards probabilistic approaches and focus more on learning rather than on human-coded knowledge. Only then will the SW scale to the real Web that human users care about. Such a turn from symbolic, crisp methods relying on hand-coded knowledge to data-driven and empirical approaches is known from the CL community. Possibly the SW is going through this turn right now in the presence of massive data, in the era of social media data, etc. Overall, one conclusion is that the Semantic Web has underestimated the complexities of natural language. One goal for the near future is to develop a roadmap by which both fields can converge, focusing on empirical and data-driven methods that can scale to the Web, to multiple languages, to the 'new' and unseen, etc. Such a convergence could follow a principle we could call *'as deep as you can, as shallow as you can get away with'*, with the goal of combining the best breed of two disciplines that have remained too orthogonal to each other in the past.

4.3 Interaction between NLP & SW Communities (Day 1)

This group concluded that the resources developed in the NLP are usually of high quality as they are maintained and curated by the community. Examples of such high-quality resources are for example WordNet, EuroWordNet, etc. In the SW community, there is a large amount of datasets in machine-readable format, such as DBpedia, Yago etc. These are typically of a lower quality as they are usually not manually validated/curated, but have been used successfully for knowledge-intensive NLP, i.e. for coreference resolution, discourse parsing, natural language generation, etc.

Overall, we notice a certain redundancy in the research carried out in both the SW and NLP communities. Problems that the SW is currently considering, i.e. named entity disambiguation using DBpedia categories, are very related to the – mainly unsupervised – WSD task that the NLP community has been working on for a number of years now. These synergies should be exploited.

We clearly need a synergistic, mutually-beneficial interaction between the NLP and Semantic web communities. In particular, we see a need for the NLP community to be convinced that the SW can support NLP tasks. There would be various ways to achieve this. On the one hand, an experimental framework could be set up to show that linked data can be exploited for NLP. Further, medium- or low-quality corpora could be created by exploiting DBpedia and linked data in order to develop approaches that can work with such a level of noise. This is in fact what is currently done in relation extraction, where an automatic alignment with DBpedia facts is exploited as 'weak' or 'distant supervision'.

4.4 Language Resources for the Semantic Web and vice versa (Day 2)

This discussion group mainly revolved around the question of how standards can be developed that facilitate the open sharing and interoperability of corpora and other language resources.

To date there are several independent initiatives developing annotation standards, including:

- ISO TC37 SC4 LAF/GrAF (LAF)
- ISO TC37 SC4 WG1 new work item on exchange protocols (EP)
- W3C Group on OpenAnnotation (OA)
- NIF and POWLA

While there are some coordination efforts between these groups, a stronger presence of the CL community in the W3C group on OpenAnnotation is desireable, implying coordinated effort and action here. A set of actions that could help to foster consensus with the goal of developing an annotation model across communities could be the following:

- Have one proposal from the CL to the SW community and vice versa, avoiding fragmentation within a single community, thus facilitating the process of consensus building.
- Work in a bottom-up and implementation-driven fashion, working from implementations to the worked-out model; rely for now on existing implementations (OLiA, MASC, etc.).
- Coach/instruct the other communities to avoid redundant work and re-discovering the wheel.
- Require availability of several implementations for each standard.

In order to convince the NLP community about the value of Linked Data, there are several options:

- Organize shared tasks across communities
- Establish an EAGLES-like effort
- Provide popular LRs in RDF for everyone to use (Babelnet, MASC, OANC, etc.)

In the short term, to increase visibility of Linked Data within the NLP community, it would be good to release as many LRs as Linked Data as possible, but also link together various resources across types (corpora, lexicons, ontologies encyclopedias, ...). In the longer term, the community should be engaged in collaborative development.

In this context, promoting openness of data seems to be the key. The EC should insist on funded projects making their linguistic resources at least freely available.

4.5 How to Improve Linked Data by the Addition of LRs (Day 2)

This group also discussed how standards in the area of LRs and Linked data should emerge. The observation was that standards should emerge as a community effort, and become de facto standards rather than just being imposed from above. In order to foster the adoption of a standard, the following is thinkable:

- Create incentives for people to adopt a standard, e.g. tools, linking into a whole ecosystem, applications
- Develop tools that allow to validate data, transform data into other formats, etc.

It was also discussed that RDF as a datamodel is particularly suited due to the following aspects:

- RDF facilitates the task of integrating and combining various vocabularies
- It supports openness of categories, allowing people to add their own linguistic categories (e.g. POS tags)

4.6 Parallel Corpora (Day 2)

One of the questions that this group discussed was to what extent multilingual lexicons derived from parallel corpora could be helpful for the Semantic Web. It was observed during the group discussion that multilingual lexicons can be extracted quickly from parallel corpora at reasonable precision. Here, precision would be certainly more important than recall. However, so far parallel corpora have not received wide attention in the SW:

It was then discussed where we could obtain parallel corpora. The Europarl and JRC corpora as well as the EU Journal and UN documents are good sources, but they will not be sufficient. In general, it is quite a challenge to extract parallel data from the Web. Google for instance uses books. The ACCURAT project has failed to deliver larger volumes of parallel data from comparable corpora. Possibly crowd-sourced translation can be an option here.

Another matter that prevents usage of corpora are licensing issues. Wikipedia is a large and multilingual as well as free resource. Newspaper data is easier to obtain now compared to 10 years ago. Unfortunately, corpora from national language institutions are not accessible.

4.7 Under-resourced Languages and Diversity (Day 3)

The group started discussing what counts as an under-ressourced language. For a language not to count as under-resourced, it should have a reasonable number of lexicons, (parallel) corpora as well as a large Wikipedia, a WordNet and a morphosyntactic analyzer. It was observed that there are many under-ressourced languages in both India and South Africa as developing countries. Some of these languages, e.g. Hindi, have huge speaker populations. In general, these languages differ considerably in the average education and economical levels of their speakers.

The situation in India is characterized by the fact that there are many languages with many different scripts. India has a strong tradition in the study of languages and is technologically ahead in many respects. WordNets and morphosyntactic resources have been developed for a number of Indian languages. However, no large national parallel corpus is available so far. To some extent, this is due to political impediments.

The general situation in South Africa is that there are generally only few language resources. NLP is not widely taught in the country and there are only scattered efforts towards the creation of language resources that are not well connected and coordinated.

Concerning the Semantic Web, the group noted that there is a lack of compelling non-first-world-elite use cases that create a clear case for coping with diversity in language and format. The group agreed that simple SW-supported use cases based on simple mobile phone interactions are needed for the SW to have an impact. One would need to engage in a dialogue with end users to design such use cases. Domains that are particularly suited to demonstrate the capability of the SW to handle heterogeneity in formats and languages would be government, health and agriculture.

4.8 The Telic Semantic Web, Cross-cultural Synchronization and Interoperability (Day 3)

This group discussed the following three topics:

- The Telic Semantic Web
- Cross-cultural synchronization
- Interoperability

The group stressed the importance of extending the current ontological focus of the SW to incorporating events, goals and emotions, as users care about causality and telicity. It was noted that it is inherently easier to specify such dimensions for artifacts rather than for natural categories.

The group also emphasized the need to develop approaches to help synchronize the infoboxes in Wikipedias for different languages. The main problem is that Wikipedias for different languages use different templates. Wikidata might solve some of these problems.

Antoine Isaac pointed out that in concrete applications (in particular in the case for Europeana), standards for multilingual search are missing. The goal would be to allow users to search for content in their language but receive content in other languages.

The group emphasized that standards alone are not enough for interoperability. For example, there is a Dictionary Server Protocol (RFC 2229) building on the IANA URI scheme. This protocol was implemented on the basis of curl. However, the protocol did not find wider adoption, which shows that standards alone will not solve the interoperability problem. But what else is needed?

4.9 Collaboration, Cross-domain Adaptation of Terminologies and Ontologies (Day 4)

This group had a vivid discussion on how to involve communities in sharing, using and developing standardized resources (terminologies, ontologies, etc.) and how to organize feedback. The careful and collaborative creation and maintenance of authoritative resources in mandated domain expert groups (e.g. ISO) was not seen as being in contradiction with the wide inclusion of communities that contribute to the standard. A promising way to include such communities is by the adoption of a combined top-down and bottom-up strategy in which distributional analysis, usage, text and term mining or crowd-sourcing can be used to enrich and extend existing authoritative resources quickly and dynamically. In general, it was agreed that there are high-quality resources (terminologies and ontologies) produced through standardization process that should be transformed into MLSW and Linked Data resources (RDF, OWL, SKOS, lemon, etc).

It was seen as very important to close the gap between the Semantic Web and existing terminological standards that exist outside of the SW. A big question is certainly how to express multilinguality properly in the SW. Extensions of SKOS by considering existing models such as ISOCAT, TBX/RDF and lemon might be required. In general, the SKOSification of existing high-quality terminologies, thesauri etc. is a goal to strive for.

Overall, the group felt that W3C standards should be more language-aware in order to facilitate the internationalization and subsequent localization of terminologies and ontologies to increase language coverage. Pervasive multilinguality requires an encompassing strategy that all involved communities should subscribe to.

The conclusions of the group's discussions where the following:

- Collaborative workflows require cross-cultural awareness, trust and openness and appropriate contexts/motivations
- We need a bootstrapping approach combining bottom-up and top-down approaches to the collaborative creation and dynamic and usage-based extension of resources, terminologies, ontologies, etc.
- We need to adapt/transform existing resources into SW standards and workflows
- Pervasive multilinguality in SW requires language-aware standards and inclusion of lemon, etc.
- Pervasive multimodality for AAL, mobile applications, etc.
- Linking the non-lingual web of content to MSW

4.10 Multilingual Web Sites (Day 4)

Nowadays, the main problem with multilingual web sites is that a user needs to visit an English site first and select from the available languages in order to get the version of a page in a particular language.

From a developer's point of view, the language functionality needs to be incorporated into the Web application that allows to select a version of a page in a particular language, so the burden is certainly on the application developer and the user. Further, it is a problem that each site can implement the multilingual support differently, which makes it more costly to implement a multilingual website and creates an idiosyncratic situation in which every website implements the multilingual access differently, which is again a burden for end users. We should aim for a solution based on a standard protocol for requesting pages in different

languages instead of having application-specific and idiosyncratic applications. This protocol should be based on transparent content negotiation.

Concluding, we need three standards:

- one for multilingual web sites,
- one for multilingual linked data, and
- one for multilingual URIs.

4.11 Use Cases for High-Quality Machine Translation (Day 4)

This group discussed potential applications and use cases for high-quality machine translation in which the input is restricted. The rationale for this is that by restricting coverage, quality can be increased. Further, by restricting the input, an ontology that serves as interlingua can be used. This is the idea pursued in the Molto project. The question to be discussed was, which applications that build on high-quality machine translation are attractive for end users. We dealt with the case of translating restaurant menues into different languages. The language and the domain are quite restricted. We could easily extract an ontology of ingredients, dishes and their translations from free resources such as Wikipedia. A use case could be one where a user is abroad and wishes to find a restaurant that serves John Dory. Wikipedia could tell us that John Dory is also called Peter's fish as well as Petersfisch or Heringskönig in German as well as saint-pierre or dorée in French. Building on such multilingual lexica, high-quality translation rules for restaurant menues could be designed or learned from examples and materialized in the form of a translation grammar that can be used by different restaurants for the translation of their menues.

4.12 Scalability to Languages, User Generated Content, Tasks (Day 4)

The group discussed the following dimensions of scalability:

- Many languages
- Different types of discourse
- Text analysis tasks

Many languages. Projection of training data across languages is a promising technique to scale to many languages. The interesting question is what requirements need to be fullfilled for the language projection to work well. Whether it works well or not depends on the properties of the language and the availability of the preprocessing components. The question is whether projection would be more effective if projection is done on the semantic level. In order to scale to many languages exploiting projection-based methods, a careful analysis of the cases in which projection breaks down is crucial and useful.

Further, one-size-fits-all, unsupervised and very simple methods (Google MT is a good example here) have come a long way, but are clearly reaching their limits. While anthropological basics seem to be more or less universal and stable across cultures, languages have certainly a different conceptual structure, and translation equivalence does not mean that concepts are the same. The question is how to design a shared conceptual structure and scale from the concept to a linguistic unit.

Different types of discourse. Scaling to new types of discourse and user-generated content is regarded as crucial. Solutions to open questions and problems are for example

published earlier in fora than by companies. Companies thus have a big interest in mining such fora. BLP has so far invested in normalization of content by training systems on such type of data. For example, MT systems have been trained on Tweets.

Text Analysis Tasks. It was discussed that current text analytics framework (UIMA) mainly support pipelines. However, for some tasks processing needs to be parallel or simultaneous.

Concerning scaling up projection across languages to different tasks, the basic assumption is that a linguistic property is stable across languages. We need to find tasks and these properties in the future.

Relation between SW and NLP. Concerning the relation between the Semantic Web and NLP, the group clearly regarded the fact that resources published as LOD have the potential of gaining more visibility. However, the SW community should provide clear use cases to other communities to motivate them for collaboration. There is a clear need for a powerful infrastructure that allows to query language resources. The SW can clearly contribute to this with their expertise on data management.

NLP can certainly support the SW in inferring links between concepts to bootstrap (see the point above on the translingual Web). The SW can support NLP in building an infrastructure for querying language resources. From the NLP perspective, the Semantic Web is an additional layer put on top of the result of a processing pipeline.

5 Talk by Kimmo Rossi

In his talk, Kimmo Rossi, Project Officer from the European Commission, briefly discussed the current restructuring going on at the European Commission. He mentioned that the old INFSO Units E.1 (Language Technology), E.2 (Data) and E.4 (PSI) would be merged into one unit G.3 called “Data value chain”. The “language technology” portfolio will continue to exist, but under the broader “Data” context. In Kimmo’s view, this provides new opportunities to explore synergies between the “Data" challenge, Semantic Web, and language technologies.

Besides discussing the current open calls and their foci, Kimmo presented the European Commission as an important producer and consumer of linked data. For example, the publications office of the EU maintains EU legislation and case law document bases. It produces online journals in 23 languages with 1 million pages/year and 6000 titles of publications. It will soon provide all its content in linked open data, free for re-use, in XML (CELLAR project). It relies on standardized metadata, ontologies and vocabularies (e.g. EuroVoc) — RDF, SKOS. EUROSTAT publishes EU statistics as 5-star linked data[56] and is a pilot provider of content to the EC Open Data Portal. The Joint Research Centre (JRC) hosts the European Media Monitor (EMM), which performs news aggregation, topical clustering and analysis at a rate of 150.000 items/day in 60 languages.

At the same time, the European Commission is a promoter of language resources, having funded the creation of language resources in over 45 projects and clusters of excellence including MetaNet. On the long-term, the goal is to build a permanent infrastructure of language resources, tools and services.

Further, several pan-European public service/administration initiatives are ongoing, such as

[56] http://eurostat.linked-statistics.org/

- EU Open Data Portal, building on Linked Data as a design principle
- epSOS (exchange of patient information, ePrescriptions, use of medical/pharma nomenclatures/classifications)
- e-CODEX (access to legal means across borders, e.g business registers, civil registers, judicial procedures)
- Peppol (pan-European Procurement platform)
- STORK (electronic identification across borders)

Data linking issues (across languages) are likely to emerge in all of them, in varying degrees and flavours.

At the end of his talk, Kimmo raised a number of administrative, cost-related and educational issues that need to be clarified in order to foster the wider adoption of Linked Data. He concluded his talk by observing that the Semantic Web will not materialize as long as it is incompatible with humans. He claimed that humans will never adopt RDF as a lingua franca and that human/natural language will prevail on the Web.

6 Non-academic Session

The non-academic session featured three presenters: Jeroen van Grondelle (BeInformed), Christian Lieske (SAP) and Felix Sasaki (W3C / DFKI). Jeroen van Grondelle presented his view on *conceptually-scope language technology*. He advocated a modular approach to HLT and distinguished three axes along which HLT can scale: i) different conceptualizations, ii) different tasks and iii) different languages. According to van Grondelle's view, we should concentrate on developing islands of HLT for particular domain conceptualizations and tasks, but following the same principles, models and frameworks so that HLT islands can be composed with other HLT islands to yield a growing ecosystem of HLT islands that constantly grow in terms of domain, task and language coverage. According to van Grondelle, this might be a positive way of lowering expectations on HLT technology.

Christian Lieske and Felix Sasaki discussed a number of practical challenges that need to be solved to making the multilingual Semantic Web a reality. They presented ten technical issues that are widely perceived as solved, arguing that they are actually not. The ten common mis-perceptions were the following:

1. *The perception is that the web stack is mature enough for a multilingual Semantic Web:* The reality is that e.g. datatypes in HTML5 currently do not cover all linguistic/cultural requirements properly. For example, data types for e-mails in Web forms only allow post-Punycode (no native "characters").
2. *It is widely believed that a focus on linguistic analysis is appropriate:* However, analytic approaches may fail due to contradictory facts. For example, the population of Amsterdam in the English and German Wikipedia differ. Thus, a generative approach – generating language from a language-neutral fact – might be worth investigation (cf. thoughts behind WikiData).
3. *It is widely believed that HTTP URIs are a suitable universal approach to identifiers:* However, for some applications other identifier schemes have to be taken into account. Examples: IRI, XRI, DOI, OID (used in the German ICD10 diagnosis catalogue).
4. *The Unicode consortium and the World Wide Web consortium host and provide the only relevant WWW standards:* Also organizations such as IETF and ISO work on Web-relevant technical standards; entities such as OASIS and HL7 are important for, amongst

others, domain-specific standards. For instance, HL7 has defined the Clinical Document Architecture (encoding, structure and semantics of clinical documents for exchange).

5. *It is widely believed that we can search already for all relevant information in one language:* However, ordinary users hardly can search for anything but textual content.
6. *It is widely perceived that anyone can already contribute to and use the Semantic Web:* However, contribution and use still require a lot of know-how and special tooling. For example, one may need Web developers to set up and use SPARQL endpoints for the twenty+ official languages in India.
7. *It is widely held that Linked Open Data is all we need:* However, without provenance information and trust, the data possibly cannot be meaningfully used. For instance, one would not want to use a non- trustworthy source to map between drug names.
8. *Supply (Research & Development) and Demand (Enterprises, Users) are connected:* Often, there is a gap between what is available, and what is known and needed.
9. *You need nothing but funding:* Not true, one needs to plan wisely, and to establish proper networks/liaisons. Without a suitable "architecture" (accompanied by education and outreach) you may not even solve the challenges related to language identifiers.
10. *HTML5 capable browsers implement language-related dimensions properly:* The reality is that some implementations even prevent certain characters from being submitted via forms.

7 Demonstration Session

In the demonstration session we had three demonstrations:

- Aarne Ranta presented a demo on the Grammatical Framework;
- Thierry Declerck presented a demo on his term analysis and term extraction tools
- Roberto Navigli presented a demo on BabelNet.
- Pushpak Bhattacharyya presented a demo on Hindi WordNet

8 Final Session

At the final slot on Friday morning we decided to carry out a discussion on a number of topics to be agreed upon during the session. The way we structured this session is by collecting a number of topics of importance and then voting on these. The five topics that emerged most prominently were the following:

- MLW standards (14 votes)
- Scalability to languages, domains and tasks (8 votes)
- Uses cases for precision-oriented HLT / MT (7 votes)
- Under-resourced languages (7)
- Lexicalization of linked data (6 votes)

We structured the discussion on each of these topics by identifying a roadmap along the following three categories (see Figure 4):

- **low-hanging fruits:** things that can be done right now

- **emerging directions:** trends and directions/problems that are emerging now and that we do not yet know how to solve or address; we will have to work on solutions to those in the mid-term
- **vision:** What expectations do we have on our technology in the long-term? What challenges do we want to solve in the future? How does this influence our developments today?

Figure 4 Snapshot of the interaction during the collection of topics, and of the state of the board at the end of the session.

We summarize the main points of discussion for each topic and for each of the three categories mentioned above:

8.1 MLW standards

A picture showing the state of the board with the post-its for standards can be found in Figure 5.

Low-hanging fruits:

- Standardized Vocabularies
  - POS tagset for all European languages (one tagset to rule them all)
  - Standard RDF models for ISO standards
  - RDF+ISOcat+CEN
  - Transparent Content Negotiation
  - Round-tripping
  - Cross-lingual link metadata
- Surveying
  - Gather a list of areas for standardization/Survey of existing standards
  - Single place for RDF vocabularies
- Syntactic interoperability
  - Create tools
  - Write wrappers
  - Use of standards (RDF) for non-standard resource

Figure 5 State of the board during the discussion on standards.

Emerging directions:

- shared vocabularies
- tools
- unify existing NLP
- Multilingual dataset unification
- Faster process
- Industry applications
- Standards for multilinguality
- DBPedia Ontology extend to many languages
- Semantic linking of LOD content
- Focus on small-scale?
- Complete onto-sem description for a small domain in a standard format
- Handling noise

Vision:

- Nothing – vision is bad for standards
- Holistic standards
- Modal logic reasoning
- Standard upper level ontology
- Language fully integrated into the LOD

8.2 Lexicalization of Linked Data

Low-hanging fruits:

- All linked data with > 1 language label
- Automatic lexicon generation
- Classify LOD as labelled/unlabelled
- Link LOD to LRs
- How lexical features link to ontological semantics
- Make lexicalization components available to Semantic Web community

Emerging directions:

- Portable natural language systems
- Hybrid, high precision systems
- Translate and localize all labels
- Collaborative lexicalization
- Link (unlabelled) LOD to LR

Vision:

- Coverage of all languages
- Common ontology for LOD
- NLG from linked data
- Common lexical features model
- Disambiguate lexicalized linked data

8.3 Scalability to Languages, Domains and Tasks

Low-hanging fruits:

- Large multilingual resources
- Big Parallel Corpora
- User Generated Content for all languages and domains
- Training (inter-disciplinary)
- Comparison of universities for potential students
- Multimodal corpora on the web
- "By hand" lost in the past, e.g., Altavista+Yahoo vs. Google

Emerging directions:

- Domain classification of LLOD
- A search engine for language resources
- Introduce orthogonality & modality
- Link social network to LD
- Collaborative construction of LR & KR
- Consumer appliance repair advice
- Information aggregation
- SMT+RBMT Lexica with frames and idioms

Vision:

- Fully ontologized LLOD
- Certification
- Infrastructure for LR+KR
- Multilingual Semantic Lexica
- Multimodal content

8.4 Use Cases for Precision-oriented HLT / MT

Low-hanging fruits:

- Agriculture, health, entertainment, financial, medicine, public services
- "People lose faith in low precision technology"
- "Critical Precision" (e.g., life and death in medicine)
- "Be quality-conscious"
- Translation
- Repository of meta-data of LRs
- Mobile
- High precision means building resources

Emerging directions:

- Business, Education, Emergencies
- Personal Assistants (Siri)
- Search by ideological position/ Semantic search
- Ontology accessibility
- Intelligent services
- Common ontology for LR meta-data

Vision:

- Jeopardy-style question answering
- Hybrids, e.g., HMT
- Scaling high precision models
- Domain-independent applications
- All apps are semantic, multilingual

8.5 Under-resourced Languages

Low-hanging fruits:

- Survey resources for LRLs
- Simplified language scale
- SW-aware roadmap
- Seminars for adults speaking LRLs
- Diversity training
- Publish existing resources as LOD/SW

Emerging directions:

- Europe and US should care more about LRLs
- Graceful degradation of NLP systems
- Interviews with local users
- Produce/mine parallel corpora, other resources
- Enabling community participation
- Good planning

Vision:

- Wide coverage
- LRLs not so LR
- Anyone who can talk is on the web
- Link all languages
- Language infrastructure for > 200 languages
- Incentives for low cost contribution (e.g., LRs)
- "One LR per child"
- Become involved with OLPC

9 Conclusion

There was wide agreement that this was a very fruitful seminar that contributed a further step towards bringing those communities together that are essential to render the vision of a multilingual Semantic Web true.

One desideratum that frequently emerged during the seminar was that the Semantic Web be more clear on what it can do for the NLP community. Philipp Cimiano argued that this is the wrong take. The Semantic Web essentially provides vocabularies and data management infrastructure that the NLP community can exploit to publish, reuse, integrate and query data. No more no less. Not much more should be expected (for now) from the Semantic Web. Of course, the SW community could do more to make the NLP community aware of the current possibilities, state-of-the-art technology, standards, etc.

The following action points were proposed at the end of the seminar:

- Publish a paper based on the results of this seminar, for example at the LRE Journal. This report could be a first step towards a publication resulting from this seminar.
- A book edited by Paul Buitelaar and Philipp Cimiano on the topics of this seminar to be published by Springer is planned. The call for book chapters will be send around to the participants of the seminar.
- We should carry out a survey of available standards and publish this survey in an appropriate form. We could start with standards developed in the context of CLARIN FLaReNet.
- A manifesto on the MSW would be nice to have as a joint view that can bring different communities together.
- We should have more showcases that show how technologies from the involved communities can be brought together.
- We should have an executive one-page summary of the seminar for funding agencies.

- The series of workshops on the Multilingual Semantic Web should continue. The question is in which form and for which audience and at which conferences to organize the workshop. Possibly, we should merge the series of Multilingual Semantic Web and Ontolex workshops.

As a direct outcome of the seminar, an article about the seminar at the Multilingual Magazine[57]. Thanks to Christian Lieske for this!

References

1. S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of ISWC/ASWC 2007*, LNCS 4825. Springer, 2007.
2. C.F. Baker, C.J. Fillmore, and J.B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, page 86–90. Association for Computational Linguistics, 1998.
3. C. Baldassarre, E. Daga, A. Gangemi, A. Gliozzo, A. Salvati, and G. Troiani. Semantic scout: making sense of organizational knowledge. *Knowledge Engineering and Management by the Masses*, page 272–286, 2010.
4. S. Berkovsky, T. Kuflik, and F. Ricci. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286, 2008.
5. T. Berners-Lee, J. Hendler, O. Lassila, et al. The Semantic Web. *Scientific American*, 284(5):28–37, 2001.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.
7. Christoph Boehm, Gerard de Melo, Felix Naumann, and Gerhard Weikum. LINDA: Distributed Web-of-Data-Scale Entity Matching. In *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM 2012)*, New York, NY, USA, 2012. ACM.
8. D. Bär, C. Biemann, I. Gurevych, and T. Zesch. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. page 435–440, 2012.
9. P. Buitelaar and P. Cimiano. *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. IOS Press, 2008.
10. P. Buitelaar, P. Cimiano, and B. Magnini. *Ontology learning from text: methods, evaluation and applications*, volume 123. IOS Press, 2005.
11. N. Calzolari. Towards a new generation of Language Resources in the Semantic Web vision. *Words and Intelligence II*, 36:63–105, 2007.
12. N. Calzolari, M. Monachini, and V. Quochi. Interoperability Framework: The FLaReNet action plan proposal. *Language Resources, Technology and Services in the Sharing Paradigm*, page 41–49, 2011.
13. N. Calzolari, V. Quochi, and C. Soria. The Strategic Language Resource Agenda, 2011.
14. C. Chiarcos. An ontology of linguistic annotations. In *LDV Forum*, volume 23, page 1–16, 2008.
15. C. Chiarcos, J. McCrae, P. Cimiano, and C. Fellbaum. Towards Open Data for Linguistics: Linguistic Linked Data. forthcoming.
16. B. Daille, B. Habert, C. Jacquemin, and J. Royauté. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257, 1996.

[57] http://www.multilingual.com/articleDetail.php?id=1977

17 M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing knowledge on the semantic web with watson. 329:1–10, 2007.

18 V. de Boer, P. De Leenheer, A. Bon, N.B. Gyan, C. van Aart, C. Guéret, W. Tuyp, S. Boyera, M. Allen, and H. Akkermans. RadioMarché: Distributed Voice- and Web-Interfaced Market Information Systems under Rural Conditions. In *CAiSE*, page 518–532, 2012.

19 E.W. De Luca. Aggregation and Maintenance of Multilingual Linked Data. *Semi-Automatic Ontology Development: Processes and Resources*, page 201–225, 2012.

20 E.W. De Luca, T. Plumbaum, J. Kunegis, and S. Albayrak. Multilingual ontology-based user profile enrichment. *MSW 2010*, page 41–42, 2010.

21 Gerard de Melo and Gerhard Weikum. Towards Universal Multilingual Knowledge Bases. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*.

22 L. Ding and T. Finin. Characterizing the semantic web on the web. *The Semantic Web-ISWC 2006*, page 242–257, 2006.

23 L. Ding, J. Shinavier, Z. Shangguan, and D. McGuinness. SameAs networks and beyond: analyzing deployment status and implications of owl: sameAs in linked data. page 145–160, 2010.

24 R. Eckart de Castilho and I. Gurevych. A lightweight framework for reproducible parameter sweeping in information retrieval. In *Proceedings of the 2011 workshop on Data infrastructures for supporting information retrieval evaluation*, page 7–10. ACM, 2011.

25 J. Eckle-Kohler and I. Gurevych. Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, page 550–560. Citeseer, 2012.

26 J. Eckle-Kohler, I. Gurevych, S. Hartmann, M. Matuschek, and C.M. Meyer. UBY-LMF–A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, page 275–282, 2012.

27 B. Ell, D. Vrandečić, and E. Simperl. Labels in the Web of Data. *The Semantic Web–ISWC 2011*, page 162–176, 2011.

28 M. Espinoza, A. Gómez-Pérez, and E. Mena. Enriching an ontology with multilingual information. *The Semantic Web: Research and Applications*, page 333–347, 2008.

29 S. Evert. Distributional Semantics. To appear.

30 Oliver Ferschke, Iryna Gurevych, and Marc Rittberger. FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia - Notebook for PAN at CLEF 2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF 2012 Labs and Workshop, Notebook Papers*, Sep 2012.

31 G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria, et al. Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*, 2006.

32 A. Gangemi, R. Navigli, and P. Velardi. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. *Proceedings of On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, page 820–838, 2003.

33 A. Gangemi and V. Presutti. Towards a Pattern Science for the Semantic Web. *Semantic Web*, 1(1-2):61–68, 2010.

34 J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae. Challenges for the multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71, March 2011.

35 I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C.M. Meyer, and C. Wirth. UBY–A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of EACL 2012*, page 580–590. Citeseer, 2012.

36 Iryna Gurevych and Torsten Zesch. Collective Intelligence and Language Resources: Introduction to the Special Issue on Collaboratively Constructed Language Resources. *Language Resources and Evaluation Journal - Special Issue on Collaboratively Constructed Language Resources*, Available online. To be printed in fall 2012., Mar 2012.

37 D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and M. von Wilamowitz-Moellendorff. Gumo–the general user model ontology. *User modeling*, page 428–432, 2005.

38 G. Hirst. The future of text-meaning in computational linguistics. In *Proceedings, 11th International Conference on Text, Speech and Dialogue*, page 3–11. Springer, 2008.

39 G. Hirst. Ontology and the Lexicon. *Handbook on ontologies*, page 269–292, 2009.

40 G. Hirst and M. Ryan. Mixed-depth representations for natural language text. *Text-Based Intelligent Systems. Lawrence Erlbaum Associates*, page 59–82, 1992.

41 J. Hoffart, F. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. page 229–232, 2011.

42 N. Ide, C. Fellbaum, C. Baker, and R. Passonneau. The manually annotated sub-corpus: a community resource for and by the people. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 68–73. Association for Computational Linguistics, 2010.

43 N. Ide and J. Pustejovsky. What does interoperability mean, any- way? Toward an operational definition of interoperability . 2010.

44 N. Ide and K. Suderman. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, page 1–8. Association for Computational Linguistics, 2007.

45 T. Kaefer, J. Umbrich, A. Hogan, and A. Polleres. Towards a Dynamic Linked Data Observatory.

46 M. Kay. The proper place of men and machines in language translation. *Machine Translation*, 12(1):3–23, 1997.

47 Mitesh Khapra, Salil Joshi, and Pushpak Bhattacharyya. It takes two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization. November 2011.

48 M.M. Khapra, S. Joshi, A. Chatterjee, and P. Bhattacharyya. Together we can: Bilingual bootstrapping for WSD. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, page 561–569. Association for Computational Linguistics, June 2011.

49 M.M. Khapra, S. Shah, P. Kedia, and P. Bhattacharyya. Projecting parameters for multilingual word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, page 459–467. Association for Computational Linguistics, August 2009.

50 M.M. Khapra, S. Sohoney, A. Kulkarni, and P. Bhattacharyya. Value for money: Balancing annotation effort, lexicon building and accuracy for multilingual wsd. In *Proceedings of the 23rd International Conference on Computational Linguistics*, page 555–563. Association for Computational Linguistics, August 2010.

51 J. Kim and I. Gurevych. UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery. In *Proceedings of the 9th NTCIR Workshop Meeting*, volume 9, page 487–494, 2011.

52 A. Klementiev and D. Roth. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. page 817–824, 2006.

53 P. Koehn. *Statistical machine translation*, volume 11. Cambridge University Press, 2010.

54 C. Mader, B. Haslhofer, and A. Isaac. Finding Quality Issues in SKOS Vocabularies. *Theory and Practice of Digital Libraries*, page 222–233, 2012.

55 J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner, et al. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, page 1–19, 2012.

56 J. McCrae, M. Espinoza, E. Montiel-Ponsoda, G. Aguado-de-Cea, and P. Cimiano. Combining statistical and semantic approaches to the translation of ontologies and taxonomies. page 116, 2011.

57 C. Meilicke, R. Garcia-Castro, F. Freitas, W. Robert van Hage, E. Montiel-Ponsoda, R. Ribeiro de Azevedo, H. Stuckenschmidt, O. Šváb Zamazal, V. Svátek, A. Tamilin, et al. Multifarm: A benchmark for multilingual ontology matching. *Journal of Web Semantics*, 15(3):62–68, 2012.

58 C.M. Meyer and I. Gurevych. What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *Proceedings of the 5th international joint conference on natural language processing (IJCNLP)*, page 883–892, 2011.

59 E. Montiel-Ponsoda, G. Aguado-de-Cea, A. Gómez-Pérez, and W. Peters. Enriching ontologies with multilingual information. *Natural language engineering*, 17(03):283–309, 2011.

60 E. Montiel-Ponsoda, J. Gracia, G. Aguado-de-Cea, and A. Gómez-Pérez. Representing translations on the semantic web. *CEUR Workshop Proceedings*, 2011.

61 R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69, 2009.

62 R. Navigli and M. Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(4):678–692, 2010.

63 R. Navigli and S.P. Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 216–225. Association for Computational Linguistics, 2010.

64 R. Navigli and S.P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

65 E. Niemann and I. Gurevych. The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, page 205–214. Citeseer, 2011.

66 A.G. Nuzzolese, A. Gangemi, and V. Presutti. Gathering lexical linked data and knowledge patterns from framenet. In *Proceedings of the sixth international conference on Knowledge capture*, page 41–48. ACM, 2011.

67 R. Parundekar, C. Knoblock, and J. Ambite. Linking and building ontologies of linked data. page 598–614, 2010.

68 T. Plumbaum, S. Wu, E.W. De Luca, and S. Albayrak. User Modeling for the Social Semantic Web. In *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation*, 2011.

69 A. Popescul and L.H. Ungar. Statistical relational learning for link prediction. In *Proceedings of the Workshop on learning statistical models from relational data*. Citeseer, 2003.

70 A. Ranta. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Center for the Study of Language and Information, 2011.

71 A.E. Richman and P. Schone. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 1–9, 2008.

72 F. Sasaki. Question answering as question-biased term extraction: a new approach toward multilingual QA. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 215–222. Association for Computational Linguistics, 2005.

73 G. Schreiber, M. van Assem, and A. Gangemi. RDF/OWL Representation of WordNet. W3C Working Draft. *http://www.w3.org/TR/2006/WD-wordnet-rdf- 20060619/*, 2006.

74 P. Sorg and P. Cimiano. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, 2012.

75 D. Spohr, L. Hollink, and P. Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. page 665–680, 2011.

76 T. Tudorache, J. Vendetti, and N.F. Noy. Web-protege: A lightweight owl ontology editor for the web. *SDR].(Cit. on p.)*, 2008.

77 P. Vossen. EuroWordNet General Document. Version 3, 1999.

Participants

- Guadalupe Aguado de Cea
Univ. Politec. de Madrid, ES
- Dimitra Anastasiou
Universität Bremen, DE
- Pushpak Bhattacharyya
Indian Institute of Technology – Bombay, IN
- Gerhard Budin
Universität Wien, AT
- Paul Buitelaar
National University of Ireland – Galway, IE
- Nicoletta Calzolari Zamorani
CNR – Pisa, IT
- Manuel Tomas Carrasco Benitez
European Commission
Luxembourg, LU
- Christian Chiarcos
USC – Marina del Rey, US
- Key-Sun Choi
KAIST – Daejeon, KR
- Philipp Cimiano
Universität Bielefeld, DE
- Ernesto William De Luca
TU Berlin, DE
- Gerard de Melo
ICSI – Berkeley, US
- Thierry Declerck
DFKI – Saarbrücken, DE
- Bo Fu
University of Victoria, CA
- Asuncion Gomez-Perez
Univ. Politec. de Madrid, ES
- Jorge Gracia
Univ. Politec. de Madrid, ES
- Marko Grobelnik
Jozef Stefan Institute – Ljubljana, SI
- Iryna Gurevych
DIPF, DE
- Sebastian Hellmann
Universität Leipzig, DE
- Graeme Hirst
University of Toronto, CA
- Chu-Ren Huang
Hong Kong Polytechnic University, HK
- Nancy Ide
Vassar College – Poughkeepsie, US
- Antoine Isaac
Europeana, The Hague, and VU University Amsterdam, NL
- Christian Lieske
SAP AG – Walldorf, DE
- John McCrae
Universität Bielefeld, DE
- Elena Montiel-Ponsoda
Univ. Politec. de Madrid, ES
- Roberto Navigli
University of Rome "La Sapienza", IT
- Sergei Nirenburg
University of Maryland, US
- Laurette Pretorius
UNISA – Pretoria, ZA
- Aarne Ranta
Chalmers UT – Göteborg, SE
- Kimmo Rossi
European Commission
Luxembourg, LU
- Felix Sasaki
DFKI / W3C – Berlin
- Gabriele Sauberer
TermNet – Wien, AT
- Hans Uszkoreit
Universität des Saarlandes, DE
- Josef van Genabith
Dublin City University, IE
- Jeroen van Grondelle
Be Informed – Apeldoorn, NL
- Daniel Vila-Suero
Univ. Politec. de Madrid, ES
- Martin Volk
Universität Zürich, CH

Report from Dagstuhl Seminar 12363

Software Defined Networking

Edited by

Pan Hui[1] and Teemu Koponen[2]

1 **TU Berlin, DE,** ben@net.t-labs.tu-berlin.de
2 **Nicira Networks Inc. – Palo Alto, US,** koponen@nicira.com

Abstract

This report documents the talks and discussions of Dagstuhl Seminar 12363 "Software Defined Networking". The presented talks represent a spectrum of industrial and academic work as well as both technical and organizational developments surrounding Software Defined Networking (SDN). The topic of SDN has garnered significant attention over the past few years in the networking community and beyond, and indeed the term "Software Defined Networking" itself carries different meaning among different circles. A key focus of the talks and discussions presented here is to capture the essence of SDN through concrete network applications, operational experience reports, and open research problems.

Seminar 05.– 08. September, 2012 – www.dagstuhl.de/12363
1998 ACM Subject Classification C.2.1 Network Architecture and Design, C.2.3 Network Operations.
Keywords and phrases Software Defined Networking, Routing, Data centers, Network Abstractions
Digital Object Identifier 10.4230/DagRep.2.9.95
Edited in cooperation with Dan Levin – dan@net.t-labs.tu-berlin.de

1 Executive Summary

Pan Hui
Teemu Koponen

Software Defined Networks (SDN) is seen as the most promising solution to resolve the challenges in realizing sophisticated network control. SDN builds its promise on the separation of the network control functions from the network switching elements. By moving the control plane out from the network elements into stand-alone servers, the switching elements can remain simple, general-purpose, and cost-effective and at the same time the control plane can rely on design principles of distributed systems in its implementation instead of being confined to distributed routing protocols.

The purpose of the seminar was to look at the current developments in this quickly evolving problem domain and identify future research challenges. The seminar brought together researchers with different domains and backgrounds. Given the high level of interest in SDN from industry, the organizers also invited many participants from companies working with SDN related networking products and services. This mix of people resulted in fruitful discussions and interesting information exchange. The structure of the seminar took advantage of these different backgrounds by focusing on themed talks and group discussions.

Organization of the Seminar

Software-Defined Networking (SDN) continues to remain relevant both for the industry and academia and indeed this was very much reflected in the backgrounds of the seminar participants; the seminar had a balanced mix of representatives both from industry and academia.

These two very active communities, industry and academia, are pursuing SDN with different mind-sets, different solutions and different implications in mind, however. The organizers felt that the interactions had been clearly insufficient in the past: practical challenges in SDN continue to remain little known in the academia whereas the industry often remains unaware of the recent useful developments in research. To this end, the two and half day seminar was explicitly structured around this observation; the goal was to allow for fruitful interactions between the industry and academia to maximize the exchange of ideas, challenges and lessons learnt between these two communities.

The seminar discussions and talks were structured around three themes:

1. Status updates. From the very definition to the ongoing standardization work, SDN is still evolving. In these talks and discussions, we dived into the ongoing work at ONF as well as the perceived hard problems to be solved.
2. Industry use cases. In this theme the focus was on exposing the academia to the practical use cases on which industry is working.
3. Implementation. The third theme dived into the details and exposed the seminar participants to both the practical implementation issues faced as well as more theoretical observations about the system design.

For the status updates the seminar had the following talks at the first day. The talks were fairly short so enough discussion could be had between the talks:

- Teemu Koponen: Evolving SDN
- Peter Feil: ONF update
- David Meyer: Hard problems in OF/SDN
- Dirk Kutscher: Northbound interfaces

The discussions after (and during) the talks also bootstrapped the evening and its group discussions about the definition of the SDN and its use cases.

The second day started with the industry use cases.

- Peter Feil: Deutsche Telekom and SDN
- James Kempf: SDN: Definition and Use Cases
- Teemu Koponen: Network virtualization
- Cedric Westphal: SDN for content management/network-based CDN emulation/transparent caching

The rest of the day was dedicated for the implementation theme and a set of short talks were given again to spark the discussion later in the evening about the implementation aspects.

- Dan Levin: State distribution trade-offs in SDN
- Nate Foster: Frenetic
- Toby Moncaster: SDN, can we (IP)FIX it?
- Andrew Moore: S/FPGA/NetFPGA

- Jarno Rajahalme: Issues in routing and tunneling in OF and OVS
- Wolfgang Riedel: Alignment of Storage, Compute and Networking
- Anders Lindgren: Use cases of SDN in information centric mobile networks

The third day was again about the use cases but this time from the academic participants. The following short talks were given with discussions between the talks:

- Christian Rothenberg: RouteFlow
- Fernando Ramos: Secure, trustworthy, resilient SDNs
- Raimo Kantola: Customer Edge Switching
- Frank Dürr: Supporting Communication Middleware with Software-Defined Networking

Outcome of the Seminar

The seminar was well received by the participants. Among the participants there were also organizers of future SDN workshops (IRTF SDN and DIMACS SDN) who signaled the intent of building their workshops around the similar discussion-oriented structure preferred at Dagstuhl.

2 Table of Contents

3 Overview of Talks

3.1 Revisiting Routing Control Platforms with the Eyes and Muscles of Software-Defined Networking

Christian Esteve Rothenberg (CPqD – Campinas, BR)

Joint work of Esteve Rothenberg, Christian; Nascimento, Marcelo; Salvador, Marcos; Correra, Carlos; Lucena, Sidney; Raszuk, Robert

This talk addresses the question on the need and opportunities of combining IP routing protocols in OpenFlow/SDN. The talk aims at raising the debate on (i) transitioning existing networks to OpenFlow/SDN, (ii) hybrid OpenFlow/SDN approaches (i.e. integration with legacy control planes), and (iii) how OpenFlow direct FIB manipulation can help IP routing control applications and enable cost-effective architectures. The research agenda of the RouteFlow project touches prior work on centralized Routing Control Platform (RCP) and its benefits in flexible routing, enhanced security, and ISP connectivity management tasks. This talk calls for input to shape the project research agenda and discusses a number of open research challenges. In addition, we present experiences from prototyping a RouteFlow Control Platform (RFCP) that implements a single node abstraction by means of an AS-wide eBGP routing service.

3.2 Software Defined Networking: overview and use cases

James Kempf (Ericsson – San Jose, US)

This talk presents perspectives and applications of SDN in the context of the wide area network, specifically the mobile core, aggregation of carrier networks as well as within the data-center. Historically, there have been many examples of split-architecture networks enabling the separation of network control from data plane forwarding – following the general principle of policy/mechanism separation. These approaches however did not introduce sufficiently powerful abstractions to realize full convergence of the very diverse set of network and service control interfaces inherent to today's carrier networks. As carrier networks become more closely integrated with services such as IPTV, customer data hosting, and general infrastructure as a service, there is increasing demand for a unified control plane and management interface for the network. For example, an SDN approach to this problem would coordinate provisioning of storage, processing, and network resources through the same management interface. These use cases depend heavily upon finding the right network and service control abstractions and interfaces. Furthermore, they also introduce complex distributed systems problems of control state management. Nevertheless, SDN presents an opportunity toward unifying today's complex tangle of network and service control management.

3.3 Evolving SDN: My view on "what the heck is it?"

Teemu Koponen (Nicira Networks Inc. – Palo Alto, US)

Despite the growing body of research and industrial initiatives based on Software Defined Networks over the past few years, the true essence of SDN remains somehow unclear. This talk shares observations of how today's notion of SDN evolved – from early attempts to improve the manageability of the IP network control plane, to current ideas on better leveraging abstractions in network design and operation. SDN, today emerged from an approach to decouple the control from forwarding, as a mechanism to enable better control (e.g., 4D, Ethane, RCP, Sane). Consequently, OpenFlow and early controller implementations such as NOX emerged as a means to realize this decoupling. The essence of SDN goes beyond the notion of decoupling control from forwarding, however. Once decoupled, modern distributed systems approaches using modular software design principles and tools can be applied to network design and operations. This talk argues that the potential to bring network design, control, and operation into the age of modern software development and deployment, to a large extent, defines the essence of SDN.

3.4 Logically Centralized? State Distribution Trade-offs in Software Defined Networks

Daniel Levin (TU Berlin, DE)

Joint work of Levin, Daniel; Wundsam, Andreas; Heller, Brandon; Handigol, Nikhil; Feldmann, Anja

Software Defined Networks (SDN) give network designers freedom to re-factor the network control plane. One core benefit of SDN is that it enables the network control logic to be designed and operated on a global network view, as though it were a centralized application, rather than a distributed system – logically centralized. Regardless of this abstraction, control plane state and logic must inevitably be physically distributed to achieve responsiveness, reliability, and scalability goals. Consequently, we ask: "How does distributed SDN state impact the performance of a logically centralized control application?"Motivated by this question, we characterize the state exchange points in a distributed SDN control plane and identify two key state distribution trade-offs. We simulate these exchange points in the context of an existing SDN load balancer application. We evaluate the impact of inconsistent global network view on load balancer performance and compare different state management approaches. Our results suggest that SDN control state inconsistency significantly degrades performance of logically centralized control applications agnostic to the underlying state distribution.

3.5 Hard Problems in Software Defined Networks

David Meyer (CISCO Systems – Eugene, US)

This talk outlines perspectives on the hard research and implementation problems to realizing modern, reliable, robust, deterministic, and practical Software Defined Networks. These problems fall roughly into the categories of technical, social, and economic challenges. Among the technical challenges, the separation of data and control planes introduces scalability and failure-resilience questions as new signalling mechanisms and failure modes may be introduced. Distributed state management of the logically centralized control plane also brings networking further into the realm of distributed and concurrent systems. Achieving the right abstractions for network control that expose enough of the underlying distributed systems complexity to prevent abstraction inversion while still ensuring sufficient control to achieve near optimal network performance goals. Hardware implementations to realize decoupling of the forwarding and control plane face very concrete problems in terms of ASIC design optimization, TCAM space and power trade-offs, etc. Given the right abstractions for network control, reasoning systems toward behavioral correctness guarantees will pose a significant challenge to achieve, but potentially enable new levels of network behavioral determinism. SDN challenges much of the existing dogma in network design and operation, and thus poses a significant sociological challenge to the establishment. SDN challenges or at least motivates revisiting fundamental philosophical design questions: Circuits vs. hop-by-hop forwarding, centralized vs. Distributed control planes, and a shift in influence bases from NetOps to DevOps all must be addressed. Finally, one can argue that the above problems all pose economic challenges to the existing vendor and operator ecosystem. Ultimately, these problems represent a sampling of the difficult problems that SDN presents to us.

3.6 Language-Based Abstractions for SDN

Nate Foster (Cornell University – NY, US)

Joint work of Guha, Arjun; Gutz Stephen; Harrison, Rob; Monsanto, Chris; Reitblatt, Mark; Rexford, Jennifer; Schlesinger, Cole; Story, Alec; Walker, David

This talk presents examples of how modular software abstraction principles, inspired from the programming languages research domain, can be used to tackle problems in networking. This work leverages a key aspect of Software Defined Networking, the programmable forwarding plane, to enable these abstractions toward realizing more modular, portable, efficient, and understandable networks.

The first principle, component modularity, is frequently used in the software engineering and programming languages domain. Through this principle, system functionality is specified and implemented once, and then reused and composed to realize more complex functionalities. By reducing code duplication, modular composition can reduce system complexity and reduce faults, enabling larger functional modules to be built from smaller, well-understood, and well-tested components. Ideally, network control logic implementation should benefit in the same fashion from the application of modular composition. For example, the forwarding plane state needed to realize the functionality of (1) network topology discovery and (2) network

traffic statistics collection overlap to a degree – that is, both functions share a common subset of instructions. In both functionalities, information about packet arrivals must be collected at the forwarding device and then reported back to a controller. The challenge in enabling these functionalities to be implemented as modular components however lies in how the forwarding state for both functionalities is combined and installed on the devices. The NetCore Language has been developed to this end, to demonstrate how to realize component modularity. High level modular components can be compiled into low-level forwarding instructions to modify forwarding state in order to correctly realize the composition of the modular functional components.

Another example of how software abstraction principles can help ensure efficient and correctly behaving networks, arises from the problem of configuration updates. More specifically, in existing networks, it is extremely hard to reason about the behavior of a network as it transitions from one configuration state to another. Even if a certain property of the network holds before and after the configuration change has been made, e.g. reachability between a given source and destination, it is incredibly difficult to know whether that property will hold at all points through the transition. Abstractions for network update enable network operators to better reason about the behavior of the network throughout the entire transition process. The key idea is that for every packet belonging to a defined flow, that packet must be forwarded at each switch through the network according to one forwarding configuration only. This behavior is realized by keeping both the old and the new forwarding states at every forwarding device within the network. Packets are tagged with an identifier to ensure that as they traverse the network, they are forwarded at each hop according to the same policy, either the old or the new, never a mixture. Consequently, for certain network properties, it becomes possible to guarantee that if the property exists in the starting and ending configuration, that property also exists throughout the configuration transition.

3.7 Customer Edge Switching

Raimo Kantola(Aalto University – Findland)

Joint work of Beijar, Nicklas; Llorente, Jesus; Leppaaho, Petri; Pahlevan, Maryam

This talk describes research efforts toward solving customer edge traffic management and considers the potential for exploiting the mechanisms of the programmable forwarding plane of Software Defined Networks. Customer-facing networks face certain challenges in providing scalable, cost-effective connectivity, while enforcing usage and service level policies and preventing unwanted and malicious traffic from reaching end-users. Customer edge switching (CES) has been proposed as an approach to allow customer-facing network operators to realize these goals, by separating customer connected devices from the public internet and mediating, where possible, connections through application-specific gateways. In theory, by acting as an intermediary between the customer edge network and the public internet, traffic filtering, admission, name resolution and route validity can be more easily established. The mechanisms to realize these behaviors in CES share some semblance with the programmable forwarding plane of Software Defined Networks. While there is ongoing research remaining to understand what aspects of CES may be facilitated with the mechanisms of SDN, proofs of concept demonstrate that certain functionality, i.e., collaborative firewalling, can be achieved.

3.8 A Research Overview – "Things that consume my time"

Andrew Moore (University of Cambridge Computer Lab – Cambridge, UK)

This talk broadly documents ongoing group research activities. A primary focus of the group is the NetFPGA, a line-rate, flexible, open networking platform for teaching and research. The NetFPGA is a PCIexpress device that enables researchers to implement among other things, hardware accelerated routers, traffic generators, and OpenFlow packet forwarding. It is an appealing platform to try out new packet-forwarding primitives at line rate. The NetFPGA is but one sub-project of another larger undertaking at Cambridge University, the (MRC^2) or Modular Research-based Composably trustworthy Mission-oriented Resilient Clouds project. This project focuses on problems in areas of Data-center switching, distributed resilience, and energy efficiency.

3.9 An Update on the Open Networking Foundation

Peter Feil (Telekom Innovation Laboratories – Berlin, DE)

This talk provides an overview of the activities and working group structure of the Open Networking Foundation (ONF), with a specific highlight on the organizational goals. The ONF was founded as a non-profit mutual benefit corporation, to promote the development and use of SDN starting with OpenFlow. Members pool their Intellectual Property Rights and may all share within this pool. There exist around 9 current working groups in charge of tasks such as OF specification extensibility, testing interoperability of implementations, marketing and education, and northbound interfaces.

3.10 Expanding SDN Primitives for Content

Cedric Westphal (Huawei US R&D Center – CA, USA)

The current functionality supported by OpenFlow-based software defined networking (SDN) includes switching, routing, tunneling, and some basic fire walling while operating on traffic flows. However, the semantics of SDN do not allow for other operations on the traffic, nor does it allow operations at a higher granularity. In this work, we describe a method to expand the SDN framework to add other network primitives. In particular, we present a method to integrate different network elements (like cache, proxy etc). Here, we focus on storage and caching, but our method could be expanded to other functionality seamlessly. We also present a method to identify content so as to perform per-content policy, as opposed to per flow policy. We have implemented the proposed mechanisms to demonstrate its feasibility.

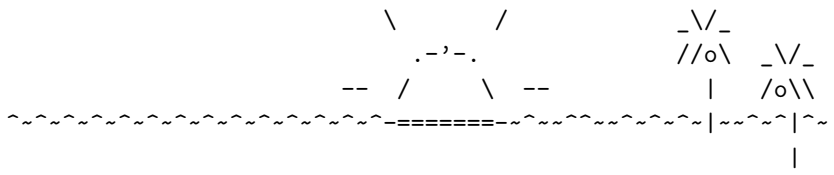
3.11 Secure, trustworthy, resilient SDNs

Fernando Ramos (University of Lisbon, FCUL, LaSIGE – Lisbon, PT)

This talk presented some of the generic ideas and preliminary research the SDN group in the University of Lisbon (www.navigators.di.fc.ul.pt) is undertaking. It's still all very early-stage, but in summary they are looking at two aspects of Software Defined Networks. First, they focus on the SDN as a distributed system. Their aim is to build secure, resilient, consistent (in its several flavours) distributed control planes. Second, they are using SDN for generic network research and to build robust middle-ware. Examples include improving fault and intrusion tolerant systems, building robust pub/sub middle-ware, connecting massive bioinformatics data processing infrastructures, increasing the efficiency of IPTV networks, among others.

4 Panel Discussions

4.1 Evening Discussion Day 1



What's our definition of SDN?

- Applying the same kind of abstractions that we've been using everywhere else in our field of computer science to the field of networking
- Breaking up the vertically integrated, monolithic black boxes to enable more flexibility and innovation based on software concepts
- Relocate control to from tightly integrated devices to "somewhere else"
- Actually, software is an irrelevant part of it. It's abstraction-defined networking and the ability to push the pieces around. The Modularization of it.
- The term software is not just the control – also the ASICs on the forwarding plane.
- The ability for consumers to have vendor-neutral APIs. The thin waist of network management is the CLI, SNMP, and the python expect script.
- Important to consider the semantics of the word "User" is the one who owns the box.

Does it really provide anything new or is it just hype?

- It is nothing new, just packaged up in a nice way that has let it gain traction
- It is a confluence of circumstance
- Routebricks disaster!
- All middle-boxes are x86. Why not everything? Because they're general and bad for everything.

What are the difficult problems ahead? What are the issues that have been definitely solved by now?

- Vendors want to keep control. (One man's opinion)

- Up to now, we don't see cheaper products.
- There's a clear tension between the cost of implementation and the performance/modularity of implementation
- We have lessons having been learned from the distributed systems in the building blocks. E.g. ONIX
- Recursion and conditionals in the data-plane
- No Consensus on the API
- Fixed depth that you can push labels
- Transformation of policy into forwarding entries
- OpenFlow has an intimate relationship with a particular hardware forwarding approach. Too hardware-coupled.
- Functional control modules have very tight couplings. E.g. in NVP a change to the configuration propagates throughout all the functional blocks

Not solved:

- Modularity: the Quantum one
- Flowvisor:
- Things are very monolithic still. Nox and Pox. Not much solved.
- It's not clear what the proper semantics of the abstractions should be
- OpenFlow will not be all things to all people.
- Is it possible to achieve complex functionality with the simpler, building blocks where complexity is moved out of the box and into the control platform. I.e., Multicast and IGMP

Got right:

- The huge miracle is the confluence of circumstances.
- The door has been jimmied open.
- OpenFlow 1.0 was very simple "so simple, even we idiots could implements" and this let the research community get started
- Low bar to entry.

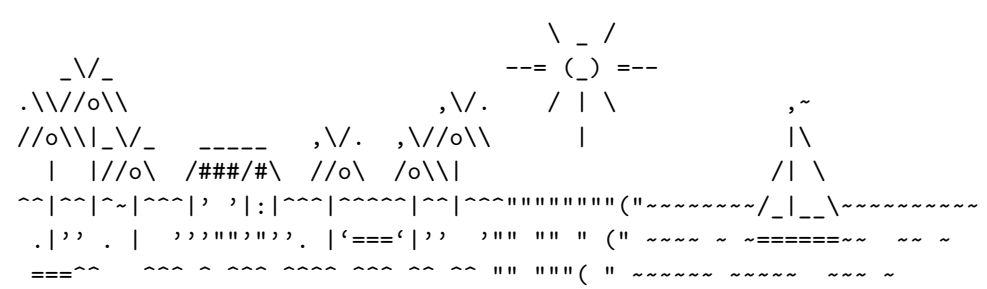
What's the role of OpenFlow in SDN, is it even needed in the end? What other interfaces should we define, why or why not?

- It's one instance of an interface to realize SDN.
- It gives us a starting point

What's the killer use case for SDN? What's the least appealing use case for it?

- The Killer use case hasn't yet been seen
- The abstraction of the network control will be the killer case for SDN

4.2 Evening Discussion Day 2



Is reactive flow processing (i.e., first packet of a flow to controller) essential to SDN? What's a flow after all?

- Reactivity is not inherent to OpenFlow. Reactivity comes from the need to handle soft-state.
- Difane
- Just think about how an IP router populates its FIB
- There's a hierarchy of life-times for pipes,

Can centralized network control ever scale?

- YES! But it depends on event rates, and where and how you centralize.
- AND it doesn't have to be physically centralized, Logically Centralize!
- We are making separate decisions about distribution of state

What does logical centralization mean?

- When you control the domain, It means you have design choices.
- Eventually Consistency
- Locking
- 2 or 3-Phase Commits
- CAP Theorem
- Transactional
- Which you can achieve depends on the network environment and workload
- It means you can aim for strong consistency models
- You try to slice the functionality of the network to keep separate the divergence
- Ben is not standing next to me right now.

Distributed systems principles are often the magic proposed for scaling, but what are really the mechanisms and principles required to scale?

- See Above
- everything stems from you need for strong consistency.

What about availability, doesn't centralization result in reduced reliability by definition?

- Logical Centralization! See above :-)

There are increasingly many language proposals for SDN. What are the fundamental problems they assist developers with?

- Network dev-ops need ways of specifying network configurations: e.g. the what of the network lookup-algorithms
- How do I specify to my network enforce the following policy and not more
- Is my network actually doing what I think its doing?
- Is the specification correct
- Bluespec
- Network debugging
- There's some real work to be done applying formal methods to reasoning about SDN

SDN is all about making network control more flexible. That tends to result in more complicated network state and systems. What are your thoughts on the implications for testing and debugging?

- By the way: What is a good metric for measuring network control flexibility?
- The total volume of configuration state goes up by orders of magnitude and it becomes too much for a human to reason about
- The abstraction makes it very hard to reason about what actually went wrong

What are the specific tools (or perhaps you have implemented one) that would be helpful?

- OFRewind (maybe. At least the idea)
- FPGAs
- vim and git (it's software!)

Participants

- Bengt Ahlgren
Swedish Institute of Computer Science – Kista, SE
- Marcus Brunner
Swisscom AG – Bern, CH
- Frank Dürr
Universität Stuttgart, DE
- Lars Eggert
NetApp Deutschland GmbH – Kirchheim, DE
- Christian Esteve Rothenberg
CPqD – Campinas, BR
- Peter Feil
Deutsche Telekom AG Laboratories, DE
- Anja Feldmann
TU Berlin, DE
- Nate Foster
Cornell University – Ithaca, US
- Howard Green
Ericsson – San Jose, US
- Pan Hui
T-labs/TU Berlin, DE
- Raimo Kantola
Aalto University, FI
- James Kempf
Ericsson – San Jose, US
- Teemu Koponen
Nicira Networks Inc. – Palo Alto, US
- Dirk Kutscher
NEC Laboratories Europe – Heidelberg, DE
- Daniel Levin
TU Berlin, DE
- Anders Lindgren
Swedish Institute of Computer Science – Kista, SE
- David Meyer
CISCO Systems – Eugene, US
- Toby Moncaster
University of Cambridge, GB
- Andrew W. Moore
University of Cambridge, GB
- Gerd Pflueger
CISCO Systems GmbH – Halbergmoos, DE
- Jarno Rajahalme
Nokia Siemens Networks – Espoo, FI
- Wolfgang Riedel
CISCO Systems GmbH – Halbergmoos, DE
- Sasu Tarkoma
University of Helsinki, FI
- Fernando Manuel Valente Ramos
University of Lisboa, PT
- Cedric Westphal
Huawei Technologies – Santa Clara, US

Report from Dagstuhl Perpectives Workshop 12371

Machine Learning Methods for Computer Security

Edited by

Anthony D. Joseph[1], Pavel Laskov[2], Fabio Roli[3], J. Doug Tygar[4], and Blaine Nelson[5]

1 **Intel Berkeley, US,** adj@eecs.berkeley.edu
2 **Universität Tübingen, DE,** pavel.laskov@uni-tuebingen.de
3 **Università di Cagliari, IT,** roli@diee.unica.it
4 **University of California – Berkeley, US,** tygar@cs.berkeley.edu
5 **Universität Tübingen, DE,** blaine.nelson@wsii.uni-tuebingen.de

Abstract

The study of learning in adversarial environments is an emerging discipline at the juncture between machine learning and computer security that raises new questions within both fields. The interest in learning-based methods for security and system design applications comes from the high degree of complexity of phenomena underlying the security and reliability of computer systems. As it becomes increasingly difficult to reach the desired properties by design alone, learning methods are being used to obtain a better understanding of various data collected from these complex systems. However, learning approaches can be co-opted or evaded by adversaries, who change to counter them. To-date, there has been limited research into learning techniques that are resilient to attacks with provable robustness guarantees making the task of designing secure learning-based systems a lucrative open research area with many challenges.

The Perspectives Workshop, "Machine Learning Methods for Computer Security" was convened to bring together interested researchers from both the computer security and machine learning communities to discuss techniques, challenges, and future research directions for secure learning and learning-based security applications. This workshop featured twenty-two invited talks from leading researchers within the secure learning community covering topics in adversarial learning, game-theoretic learning, collective classification, privacy-preserving learning, security evaluation metrics, digital forensics, authorship identification, adversarial advertisement detection, learning for offensive security, and data sanitization. The workshop also featured workgroup sessions organized into three topic: machine learning for computer security, secure learning, and future applications of secure learning.

Seminar 09.–14. September, 2012 – www.dagstuhl.de/12371

1998 ACM Subject Classification C.2.0 Computer-Communication Networks (General): Security and Protection (e.g., firewalls); D.4.6 Operating Systems (Security and Protection); I.2.6 Artificial Intelligence (Learning); I.2.7 Artificial Intelligence (Natural Language Processing); I.2.8 Artificial Intelligence (Problem Solving, Control Methods, and Search); K.4.1 Computers and Society (Public Policy Issues): Privacy; K.6.5 Management of Computing and Information Systems (Security and Protection)

Keywords and phrases Adversarial Learning, Computer Security, Robust Statistical Learning, Online Learning with Experts, Game Theory, Learning Theory

Digital Object Identifier 10.4230/DagRep.2.9.109

Machine Learning Methods for Computer Security, *Dagstuhl Reports*, Vol. 2, Issue 9, pp. 109–130
Editors: Anthony D. Joseph, Pavel Laskov, Fabio Roli, J. Doug Tygar, and Blaine Nelson

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

1 Executive Summary

Anthony D. Joseph
Pavel Laskov
Blaine Nelson
Fabio Roli
Doug Tygar

Arising organically from a variety of independent research projects in both computer security and machine learning, the topic of machine learning methods for computer security is emerging as a major direction of research that offers new challenges to both communities. Learning approaches are particularly advantageous for security applications designed to counter sophisticated and evolving adversaries because they are designed to cope with large data tasks that are too complex for hand-crafted solutions or need to dynamically evolve. However, in adversarial settings, the assets of learning can potentially be subverted by malicious manipulation of the learner's environment. This exposes applications that use learning techniques to a new type of security vulnerability in which an adversary can adapt to counter learning-based methods. Thus, unlike most application domains, computer security applications present a unique data domain that requires careful consideration of its adversarial nature to provide adequate learning-based solutions—a challenge requiring novel learning methods and domain-specific application design and analysis. The Perspectives Workshop, "Machine Learning Methods for Computer Security", brought together prominent researchers from the computer security and machine learning communities interested in furthering the state-of-the-art for this fusion research to discuss open problems, foster new research directions, and promote further collaboration between the two communities.

This workshop focused on tasks in three main topics: the role of learning in computer security applications, the paradigm of secure learning, and the future applications for secure learning. In the first group, participants discussed the current usage of learning approaches by security practitioners. The second group focused of the current approaches and challenges for learning in security-sensitive adversarial domains. Finally, the third group sought to identify future application domains, which would benefit from secure learning technologies.

Within this emerging field several recurrent themes arose throughout the workshop. A major concern that was discussed throughout the workshop was an uneasiness with machine learning and a reluctance to use learning within security applications and, to address this problem, participants identified the need for learning methods to provide better transparency, interpretability, and trust. Further, many workshop attendees raised the issue of how human operators could be incorporated into the learning process to guide it, interpret its results, and prevent unintended consequences, thus reinforcing the need for transparency and interpretability of these methods. On the learning side, researchers discussed how an adversary should be properly incorporated into a learning framework and how the algorithms can be designed in a game-theoretic manner to provide security guarantees. Finally, participants also identified the need for a proper characterization of a security objective for learning and for benchmarks for assessing an algorithm's security.

This document summarizes the presentations and working groups held at the 2012 "Machine Learning Methods for Computer Security" Dagstuhl Perspectives Workshop. Sections 3, 4 and 5 summarize the invited presentations held by the workshop's participants. Section 6 then provides a short summary of the topics discussed by each of the workshop's

three workgroups. Finally, the open problems discussed during the workshop are summarized in Section 7 and are followed by our acknowledgments and a list of the workshops attendees.

2 Contents

3 Talks I: Overview of Adversarial Machine Learning

3.1 Adversarial Attacks Against Machine Learning and Effective Defenses

Richard P. Lippmann (MIT – Lexington, US)

Machine learning is widely used to solve difficult security problems by adaptively training on large databases. Examples include computer spam detection, antivirus software, computer intrusion detection, automated Internet search engines such as Google, credit-card fraud detection, talker identification by voice, and video surveillance. Many of these systems face active adversaries with strong financial incentives to defeat accurate performance. Just as humans are susceptible to fraud and misdirection, many of these new learning systems are susceptible to adversarial attacks. This presentation provides a taxonomy of the types of adversarial attacks that can be launched against learning systems and also a summary of effective defenses that can be used to counter these attacks. This analysis is meant to raise the awareness of weaknesses in many widely deployed learning systems, of successful defenses to counter adversarial attacks, and of the arms race this interaction engenders.

3.2 Adversarial Machine Learning, Part I

Doug Tygar (University of California – Berkeley, US)

Joint work of Tygar, J. D.; Barreno, Marco; Huang, Ling; Joseph, Anthony D.; Nelson, Blaine; Rao, Satish; Rubinstein, Benjamin I. P.
Main reference L. Huang, A.D. Joseph, B. Nelson, B.I.P. Rubinstein, J.D. Tygar, "Adversarial machine learning," in Proc. of the 4th ACM Workshop on Security and Artificial Intelligence (AISec'11), pp. 43–58, ACM.
URL http://dx.doi.org/10.1145/2046684.2046692

This talk is the first part of a two-part presentation on research on adversarial machine learning research at UC Berkeley: the study of effective machine learning techniques against an adversarial opponent. In this talk, we give a taxonomy for classifying attacks against online machine learning algorithms; discuss application-specific factors that limit an adversary's capabilities; introduce two models for modeling an adversary's capabilities; explore the limits of an adversary's knowledge about the algorithm, feature space, training, and input data; explore vulnerabilities in machine learning algorithms; discuss countermeasures against attacks; introduce the evasion challenge; and discuss privacy-preserving learning techniques. This talk focuses particularly on the taxonomy for classifying attacks and the technology of Reject on Negative Impact.

3.3 Adversarial Machine Learning, Part II

Anthony D. Joseph (University of California – Berkeley, US)

Joint work of Marco Barreno, Ling Huang, Alex Kantchelian, Blaine Nelson, Justin Ma, Satish Rao, Benjamin I.P. Rubinstein, and J. Doug Tygar

This talk is the second part of a two-part presentation on research on adversarial machine learning research at UC Berkeley: the study of effective machine learning techniques against an adversarial opponent. In this talk, we discuss attacks against a network anomaly detector and an approach to making the detector more robust against such attacks; present approaches to the near-optimal evasion problem for convex-inducing classifiers; and introduce the problem of spam detection for social networks and a content complexity-based approach to detecting spam in such environments.

3.4 Pattern recognition systems under attack: issues learned in Cagliari

Fabio Roli (Università di Cagliari, IT)

In this talk, I discuss the research experience of the PRA Lab (prag.diee.unica.it) of the University of Cagliari on adversarial pattern recognition. I start with our early work, at the end of the 1990s, on multiple classifiers for intrusion detection in computer networks, then I move to our work on evade-hard multiple classifiers and the exciting digression on image-spam filtering, up to our current activity on countermeasures against spoofing attacks in biometric systems. A few issues that I learned in the context of real applications are the focus of my talk. I do not provide the details of the partial solutions we proposed, as the talk's goal is most of all sharing some issues that, I hope, can stimulate discussion. This talk is coordinated with the other three participants coming from my Lab.

3.5 Attack Detection in Networks and Applications: lessons learned in Cagliari

Giorgio Giacinto (Università di Cagliari, IT)

Joint work of Giacinto, Giorgio; Ariu, Davide; Corona, Igino; Roli, Fabio

The investigation of Machine Learning paradigms for detecting attacks against networked computers was a response to the weaknesses of attack signatures. As a matter of fact, signatures usually capture just some characteristics of the attack, thus leaving room for the attacker to produce the same effects by applying slight variations in the way the attack is crafted.

The generalization capability of machine learning algorithms has encouraged many researchers to investigate the possibility of detecting variations of known attacks. While machine learning succeeded in achieving this goal in a number of security scenarios, it was

also a source of large volumes of false alarms. We learned that to attain the trade-off between detection rate and false alarm rate was not only a matter of the selection of the learning paradigm, but it was largely dependent on the problem statement. Source data selection, feature definition, and model selection have to be carefully crafted to attain the best trade-off between detection accuracy, generalization capability, and false alarm generation. These issues have been outlined by referring to the detection of attacks against web applications.

References

1 I. Corona, R. Tronci, G. Giacinto, *SuStorID: A Pattern Recognition System to the Protection of Web Services*, In the Proceedings of the 21st International Conference on Pattern Recognition, Japan, Nov 11-15, 2012 (in press).

2 R. Perdisci, I. Corona, G. Giacinto, *Early Detection of Malicious Flux Networks via Large-Scale Passive DNS Traffic Analysis*, IEEE Trans. on Dependable and Secure Computing, 9(5), 2012, pp. 714–726

3 D. Ariu, R. Tronci, G. Giacinto, *HMMPayl: an Intrusion Detection System based on Hidden Markov Models*, Computers & Security, 30, 2011, pp. 221–241

4 I. Corona, G. Giacinto, C. Mazzariello, F. Roli, C. Sansone, *Information fusion for computer security: State of the art and open issues*, Information Fusion, 10, 2009, pp. 274–284

5 G. Giacinto, R. Perdisci, M. Del Rio, F. Roli, *Intrusion detection in computer networks by a modular ensemble of one-class classifiers*, Information Fusion, (1), 2008, pp. 69–82

6 G. Giacinto, F. Roli and L. Didaci, *Fusion of multiple classifiers for intrusion detection in computer networks*, Pattern Recognition Letters, 24(12), 2003, pp. 1795–1803

3.6 Security evaluation of pattern classifiers: lessons learned in Cagliari

Battista Biggio (Università di Cagliari, IT)

Joint work of Biggio, Battista; Fumera, Giorgio; Roli, Fabio

Pattern recognition systems are increasingly being used in adversarial environments like biometric authentication, network intrusion detection and spam filtering tasks, in which data can be adversely manipulated by humans to undermine the outcomes of an automatic analysis. Current pattern recognition theory and design methods do not explicitly consider the intrinsic, adversarial nature of these problems. Consequently, pattern recognition systems exhibit vulnerabilities which can be exploited by an adversary to make them ineffective. This may limit their widespread adoption as potentially useful tools in many applications. Extending pattern recognition theory and design methods to adversarial settings is thus a very relevant research direction, which has not yet been pursued in a systematic way.

In this talk I discuss a general framework that addresses one of the main open issues in the field of adversarial machine learning, namely, the security evaluation of pattern classifiers. The goal of such analysis is to give a more complete view of the classifier performance in adversarial environments, by assessing the performance degradation that may be incurred under different, potential attacks. Depending on the application, this may lead to different design choices; for instance, the selection of a different classification model, or parameter setting. Our framework is based on an explicit model of adversary and data distribution, and encompasses, in a coherent and unifying way, different ideas, models and methods proposed in the adversarial classification literature thus far. It can also be exploited to design more secure classifiers. Some application examples and research directions will also be discussed.

This talk is coordinated with the other three participants coming from the PRA Lab (prag.diee.unica.it) of the University of Cagliari.

3.7 Evade-hard multiple classifiers: lessons learned in Cagliari

Giorgio Fumera (Università di Cagliari, IT)

Joint work of Fumera, Giorgio; Biggio, Battista; Roli, Fabio

Multiple classifier systems (MCSs) have been considered for decades in the pattern recognition and machine learning fields as a technique for improving classification accuracy, with respect to the traditional approach based on the design of a single classifier. Recent theoretical results on adversarial classification, as well as tools used in real adversarial environments, like intrusion detection, spam filtering, and biometric identity recognition, lead our research group to investigate whether MCSs can also be useful to increase the "hardness of evasion". Besides providing some (partial) answers to this question, our analysis pointed out that the "hardness of evasion" of pattern classifiers must be defined and evaluated taking into account the characteristics and constraints of the specific application at hand, as well as using a proper adversary's model. This talk is coordinated with the other three participants coming from my Lab (Fabio Roli, Giorgio Giacinto and Battista Biggio).

4 Talks II: Adversarial Learning and Game-theoretic Approaches

4.1 Machine Learning in the Presence of an Adversary

Tobias Scheffer (Universität Potsdam, DE)

Joint work of Brückner, Michael; Kanzow, Christian; Scheffer, Tobias
Main reference M. Brückner, C. Kanzow, T. Scheffer, "Static Prediction Games for Adversarial Learning Problems," in Journal of Machine Learning Research 13:2617–2654, 2012.
URL http://jmlr.csail.mit.edu/papers/volume13/brueckner12a/brueckner12a.pdf

Machine learning algorithms are commonly based on the assumption that data at training and application time are governed by identical distributions. This assumption is violated when the test data are generated by an adversary in response to the presence of a predictive model. A number of robust learning models have been studied that are based on the worst-case assumption that the adversary will afflicts such changes to the data at application time that achieve the greatest possible adverse effect. This assumption, however, is in many cases overly pessimistic and does not necessarily lead to the ideal outcome if the adversary pursues a goal that is in conflict with, but not necessarily directly antagonistic to, the goal of the learner. We model this interaction as a non-zero-sum, non-cooperative game between learner and data generator. The game-theoretic framework enables us to explicitly model the players' interests, their possible actions, their level of knowledge about each other, and the order at which they commit to their actions. We first assume that both player choose their actions simultaneously, without the knowledge of their opponent's decision. We identify conditions under which this Nash prediction game has a unique Nash equilibrium, and derive algorithms that find the equilibrial prediction model. As a second case, we consider a data generator who is potentially fully informed about the move of the learner. This setting establishes a Stackelberg competition. We derive a relaxed optimization criterion to determine the solution of this game and show that this Stackelberg prediction game generalizes existing prediction models. In case studies on email spam filtering, we empirically explore properties of all derived models as well as several existing baseline methods.

References

1 Michael Brückner, Christian Kanzow, and Tobias Scheffer. *Static Prediction Games for Adversarial Learning Problems*. In the Journal of Machine Learning Research 13:2617–2654, 2012.
2 Michael Brückner and Tobias Scheffer. *Stackelberg games for adversarial prediction problems*. In KDD 2011: Proceedings of the 17th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, 2011.

4.2 Sparse reward processes and multi-task security games

Christos Dimitrakakis (EPFL – Lausanne, CH)

Joint work of Dimitrakakis, Christos; Auer, Peter
Main reference C. Dimitrakakis, "Sparse Reward Processes," arXiv:1201.2555v2 [cs.LG] , 2012.
URL http://arxiv.org/abs/1201.2555v2

In many security applications, the importance of different resources to be protected is unknown, or arbitrarily changing. In this case, the agent must automatically adapt to the new goals, while continuing to gather information about the environment. The main problem is how much should the agent focus its attention to the task at hand, and how much he should gather information about parts of the environment which may not be directly related to the current task, but which may be relevant for future tasks.

We formalize this setting by introducing a class of learning problems where the agent is presented with a series of tasks. Intuitively, if there is a relation among those tasks, then the information gained during execution of one task has value for the execution of another task. Consequently, the agent might not have to explore its environment beyond the degree necessary to solve the current task.

This paper develops a decision theoretic setting that generalizes standard reinforcement learning tasks and captures these intuitions. More precisely, we introduce sparse reward processes, as a type of multi- stage stochastic game between a learning agent and an opponent. The agent acts in an unknown environment, according to a utility that is arbitrarily selected by the opponent. In the case of security games, the opponent is in fact our client, and the utility is related to the value of our resources and the costs of protection.

Apart from formally describing the setting, we link it to bandit problems, bandits with covariates and factored MDPs. Finally, we examine the behavior of a number of learning algorithms in such a setting, both experimentally and theoretically.

4.3 Near-Optimal Node Blacklisting in Adversarial Networks

Aikaterini Mitrokotsa (EPFL – Lausanne, CH)

Joint work of Dimitrakakis, Christos; Mitrokotsa, Aikaterini
Main reference C. Dimitrakakis, A. Mitrokotsa, "Near-Optimal Node Blacklisting in Adversarial Networks," arXiv:1208.5641 [cs.CR], 2012.
URL http://arxiv.org/abs/1208.5641v1

Many communication networks contain nodes which may misbehave, thus incurring a cost to the network operator. We consider the problem of how to manage the nodes when the operator receives a payoff for every moment that a node stays within the network, but where each malicious node incurs a hidden cost. The operator only has some statistical information about each node's type, and never observes the cost. We consider the case when there are two possible actions: removing a node from a network permanently, or keeping it for at least one more time-step in order to obtain more information. Consequently, the problem can be seen as a special type of intrusion response problem, where the only available response is blacklisting. We first examine a simple algorithm (HiPER) which has provably good performance compared to an oracle that knows the type (honest or malicious) of each node. We then derive three other approximate algorithms by modeling the problem as a Markov

decision process. To the best of our knowledge, these algorithms have not been employed before in network management and intrusion response problems. Through experiments on various network conditions, we conclude that HiPER performs almost as well as the best of these approaches, while requiring significantly less computation.

4.4 Formalizing Secure Learning: Quantifying the Security of a Learning Algorithm as a Basis for Assessing and Designing Classifiers

Blaine Nelson (Universität Tübingen, DE)

Joint work of Nelson, Blaine; Biggio, Battista; Laskov, Pavel
Main reference B. Nelson, B. Biggio, P. Laskov., "Understanding the Risk Factors of Learning in Adversarial Environments," in Proc. of the 4th ACM Wworkshop on Security and Artificial Intelligence (AISec '11). pp. 87-92, ACM.
URL http://dx.doi.org/10.1145/2046684.2046698

Machine learning algorithms are rapidly emerging as a vital tool for data analysis and autonomic systems because learners can infer hidden patterns in large complicated datasets, adapt to new behaviors, and provide statistical soundness to decision-making processes. This makes them useful for many emerging tasks in security, networking, and large-scale systems applications. Unfortunately, learning techniques also expose these systems to a new class of security vulnerabilities—learners themselves can be susceptible to attacks. Many common learning algorithms were developed under the assumption that training data is from a natural or well-behaved distribution. However, these assumptions are perilous in a security sensitive setting. With financial incentives encouraging ever more sophisticated adversaries, attacks increasingly target these learners (a prime example is how spammers have adapted their messages to thwart the newest spam detectors). An intelligent adversary can alter his approach based on knowledge of the learner's shortcomings or mislead it by cleverly crafting data to corrupt the learning process.

For this reason, security analysis is a crucial element for designing a practical learning system and for providing it with a sound foundation [1, 2]. However, to properly assess a system's security, the community needs appropriate notions for measuring and benchmarking the security of a learner. Part of this task has already be accomplished: qualitative assessments of security threats have been developed and measures from areas like robust statistics [3, 4] and game-theoretic learning [5] provide a basis for assessing security. However, a comprehensive measure of a learner's security (akin to differential privacy in privacy-preserving learning) has yet to be fully developed or widely accepted as a criteria for assessing classifiers.

Measures of a learner's security are essential to gain a better understanding of the security properties of learning and, ultimately, for their successful deployment in a multitude of new domains and will form the core of a more formal approach to secure learning. In this talk, we motivate the need for security measures of learning as a basis for systematically advancing secure learning. We overview the prior work for assessing a learner's stability and what needs to be done to formalize the notion of secure learning. We also introduce a notion of adversarial corruption that is directly incorporated into a learning framework and derive from it a new criteria for classifier robustness to adversarial contamination.

References

1 Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. *The Security of Machine Learning*. Machine Learning, 81(2):121–148, 2010.

2 Pavel Laskov and Marius Kloft. *A Framework for Quantitative Security Analysis of Machine Learning*. In Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence (AISec), pages 1–4, 2009.

3 Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, New York, NY, USA, 1986.

4 Peter Huber. *Robust Statistics*. John Wiley & Sons, New York, NY, USA, 1981.

5 Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

4.5 Convex Adversarial Collective Classification

Daniel Lowd (University of Oregon, US)

Joint work of Torkamani, Mohamadali; Lowd, Daniel

Many real-world domains, such as web spam, auction fraud, and counter-terrorism, are both relational and adversarial. Previous work in adversarial machine learning has assumed that instances are independent from each other, both when manipulated by an adversary and labeled by a classifier. Relational domains violate this assumption, since object labels depend on the labels of related objects as well as their own attributes.

In this talk, I present a novel method for robustly performing collective classification in the presence of a malicious adversary that can modify up to a fixed number of binary-valued attributes. This method is formulated as a convex quadratic program that guarantees optimal weights against a worst-case adversary in polynomial time. In addition to increased robustness against active adversaries, this kind of adversarial regularization can also lead to improved generalization even when no adversary is present. In experiments on real and simulated data, our method consistently outperforms both non-adversarial and non-relational baselines.

4.6 Data Privacy and Machine Learning

Benjamin I. P. Rubinstein (Microsoft Research – Mountain View, US)

Joint work of Rubinstein, Benjamin I. P.; Bartlett, Peter; Huang, Ling; Narayanan, Arvind; Shi, Elaine; Taft, Nina

Main reference B.I.P. Rubinstein, P.L. Bartlett, L. Huang, N. Taft, "Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning," in Special Issue on Statistical and Learning-Theoretic Challenges in Data Privacy, Journal of Privacy and Confidentiality, 4(1), pp. 65–100, 2012.

URL http://repository.cmu.edu/jpc/vol4/iss1/4/

The ubiquitous need for analyzing privacy-sensitive information—including health records, personal communications, product ratings and social network data—is driving significant

interest in privacy-preserving data analysis across several research communities. This talk describes two projects related to data privacy and machine learning.

The first (theoretical) part explores the release of Support Vector Machine (SVM) classifiers while preserving the differential privacy of training data. We present efficient mechanisms for finite-dimensional feature mappings and for (potentially infinite-dimensional) mappings with translation- invariant kernels. In the latter case, our mechanism borrows a technique from large-scale learning to learn in a finite-dimensional feature space whose inner-product uniformly approximates the desired feature space inner-product (the desired kernel) with high probability. Differential privacy is established using algorithmic stability, a property used in learning theory to bound generalization error. Utility—when the private classifier is pointwise close to the non-private classifier with high probability—is proven using smoothness of regularized empirical risk minimization with respect to small perturbations to the feature mapping. We conclude the part with lower bounds on the differential privacy of any mechanism approximating the SVM.

The second (experimental) part of the talk describes a winning entry to the IJCNN 2011 Social Network Challenge run by Kaggle.com which is a crowd- sourcing platform for machine learning tasks. The goal of the contest was to promote research on real-world link prediction, and the dataset was a graph obtained by crawling the popular Flickr social photo sharing website, with user identities scrubbed. By applying de-anonymizing attacks on much of the competition test set using our own Flickr crawl, we were able to effectively game the competition. Our attack represents a new application of de-anonymization to gaming machine learning contests, suggesting changes in how future machine learning competitions should be run.

5 Talks III: Secure Learning Applications, Evaluation, and Practices

5.1 Does My Computer (Really) Need My Password to Recognize Me?

Saša Mrdović (University of Sarajevo, SEU)

Basic idea: Computers (and Web sites) we use have enough information on us to enable them to authenticate us without need for (classic) authentication information based on shared secret that needs to be remembered and kept by us and the system. User authentication is, usually, a one time process executed before each interactive session between a user and a machine (OS, Web application, ...). The user provides little piece(s) of information which confirms that she is entitled to take certain identity that system recognizes. From that moment on, the subject that has been authenticated is bound to her given identity. Authentication is complete with all the rights assigned to given identity and permanent for the duration of the session. A question is if the system should relay on authentication information only to give full user rights and for the whole duration of the session. Continuous authentication has been proposed as a mechanism that continuously re-confirms user's identity during a session. It is usually based on biometric information such as keystroke dynamics or visual data on users obtained through the camera. A limitation of user rights has been implemented through re-authentication before tasks that require elevated privileges can be executed (sudo, UAC). Forensic investigations that I have performed on computers and Web applications showed me that the mentioned systems know things about users that nobody but the user himself should know. Stored data could be used for creation of authentication information that user already knows (does not have to remember) and that is, at the same time, hard for anybody else to know. This is a version of dynamic knowledge based authentication (KBA). Machine learning would be an excellent tool to enable user authentication based on data already stored about user's actions. It could enable continuous authentication as well as authentication for the elevation of privileges. It uses existing data and does not raise (additional) privacy issues. This presentation quickly covers issues with current authentication systems, existing continuous authentication research work and knowledge-based authentication implementations. It gives initial ideas about how a working system could be implemented with its pros and cons.

5.2 Evaluating Classifiers in Adversarial Environments

Alvaro Cárdenas (Fujitsu Labs of America Inc. – Sunnyvale, US)

In machine learning, classifiers are traditionally evaluated based on a testing dataset containing examples of the negative (normal) class and the positive (attack) class. However, in adversarial environments there are many practical situations where we cannot obtain examples of the attack class a priori. There are two main reasons for this: (1) by definition, we cannot obtain examples of zero-day attacks, and (2) using attack examples which are generated independently of the classifier implicitly assumes that the attacker is not adaptive and will not try to evade our detection mechanism.

We argue that instead of using a set of attack samples for evaluating classifiers, we need to find the worst possible attack for each classifier and evaluate the classifier by considering the costs of this worst-case attack.

As a result we can obtain a new trade-off curve as an alternative to ROC curves. The x-axis is still the false positive rate, which can be computed by a dataset of negative examples (which are generally easy to obtain). However, instead of estimating the true positive rate for the y-axis (as in traditional ROC curves) we compute the "cost" of the worst undetected attack by crafting worst-case (in terms of defender cost) attacks. We use the cost of these undetected attacks in the y-axis.

A new area where these types of attacks are easy to create is in the protection of cyber-physical systems and other critical infrastructures, where the state of the system can be associated with a monetary cost. Examples of this new evaluation method are given in the context of electricity theft in the smart metering infrastructure [1] and a chemical reactor [2].

References

1 Daisuke Mashima and Alvaro A. Cardenas. *Evaluating Electricity Theft Detectors in Smart Grid Networks*. In Proceedings of the 15th International Symposium on Research in Attacks, Intrusions, and Defenses (RAID), pp 210-229. Amsterdam, The Netherlands, September 12-14, 2012.

2 Alvaro A. Cardenas and Saurabh Amin and Zong-Syun Lin and Yu-Lun Huang and Chi-Yen Huang and Shankar Sastry. *Attacks against process control systems: risk assessment, detection, and response*. In Proceedings of the 6th ACM Symposium on Information, Computer, and Communications Security (ASIACCS). pp 355-366. Hong Kong, March 22-24, 2011.

5.3 Detection of Malicious PDF Files Based on Hierarchical Document Structure

Nedim Šrndić (Universität Tübingen, DE)

Joint work of Šrndić, Nedim; Laskov, Pavel

Malicious PDF files remain a real threat, in practice, to masses of computer users even after several high-profile security incidents and various security patches issued by Adobe and other vendors. This is a result of widely-used outdated client software and of the expressiveness of the PDF format which enables attackers to evade detection with little effort. Apart from traditional antivirus products, which are always a step behind the attackers, few methods have been proposed which can be deployed for the protection of end-user systems. This talk introduces a highly performant static method for detection of malicious PDF documents which, instead of performing analysis of JavaScript or any other content for detection, relies on essential differences in the structural properties of malicious and benign PDF files. The effectiveness of the method was evaluated on a data corpus containing about 210,000 real-world malicious and benign PDF files and it outperforms each of the 43 antivirus engines at VirusTotal and other specialized detection methods. Additionally, a comparative evaluation of several learning setups with regard to resistance against adversarial evasion will be presented which shows that our method is almost completely immune to sophisticated attack scenarios.

5.4 Detecting Adversarial Advertisements in the Wild

Nathan Ratliff (Google – Pittsburgh, US)

Joint work of Sculley, D.; Otey, Matthew; Pohl, Michael; Spitznagel, Bridget; Hainsworth, John; Zhou, Yunkai
Main reference D. Sculley, M.E. Otey, M. Pohl, B. Spitznagel, J. Hainsworth, Y. Zhou, "Detecting Adversarial Advertisements in the Wild," in Proc. of the Int'l Conf. on Data Mining and Knowledge Discovery (KDD'11).
URL http://www.eecs.tufts.edu/ dsculley/papers/adversarial-ads.pdf

Online advertising is a growing multi-billion dollar industry. Google has an extensive set of ad networks, and wants to provide the best possible advertising experience for both users and advertisers. By providing relevant, high-quality ads to users, customers are more likely to trust the quality of Google's ad networks and make purchases. However, some malicious advertisers attempt to exploit this trust and use questionable means to take users' money. In this talk, we will present the work of Google's Landing Page Quality team, whose purpose it is to remove these adversarial advertisers from Google's ad network. We will give an overview of the systems and methods we use to detect and remove such adversarial advertisers.

References

1 D. Sculley, Matthew Eric Otey, Michael Pohl, Bridget Spitznagel, John Hainsworth, and Yunkai Zhou. *Detecting Adversarial Advertisements in the Wild*. In KDD 2011: Proceedings of the 17th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery. August, 2011.

5.5 Deceiving Authorship Detection

Rachel Greenstadt (Drexel University – Philadelphia, US)

Joint work of Greenstadt, Rachel; Afroz, Sadia; Brennan, Michael; McDonald, Andrew; Caliskan, Aylin; Stolerman, Ariel
Main reference S. Afroz, M. Brennan, R. Greenstadt, "Detecting Hoaxes, Frauds, and Deception in Writing Style Online," IEEE Security and Privacy, 2012
URL https://www.cs.drexel.edu/ sa499/papers/oakland-deception.pdf

In digital forensics, questions often arise about the authors of documents: their identity, demographic background, and whether they can be linked to other documents. The field of stylometry uses linguistic features and machine learning techniques to answer these questions. While stylometry techniques can identify authors with high accuracy in non-adversarial scenarios, their accuracy is reduced to random guessing when faced with authors who intentionally obfuscate their writing style or attempt to imitate that of another author.

In this talk, I will discuss my lab's work in the emerging field of adversarial stylometry. We will discuss our results detecting deception in writing style that may indicate a modified document, demonstrating up to 86% accuracy in detecting the presence of deceptive writing styles. I will also discuss our efforts to aid individuals in obfuscating their writing style in order to maintain anonymity against multiple forms of machine learning based authorship recognition techniques and end with some work in progress extending our research to multilingual data sets and investigating stylometry as a means of authentication.

5.6 Vulnerability Extrapolation: Machine Learning and Offensive Security

Konrad Rieck (Universität Göttingen, DE)

Joint work of Yamaguchi, Fabian; Lottmann, Markus; Rieck, Konrad
Main reference Fabian Yamaguchi, Markus Lottmann, and Konrad Rieck. Generalized Vulnerability Extrapolation using Abstract Syntax Trees. Annual Computer Security Applications Conference (ACSAC), December 2012

Machine learning has been traditionally used as a defensive tool in computer security. Only a little research has studied whether and how learning methods can be applied in an offensive setting. In this talk, we explore this research direction and present a learning-based approach for discovering vulnerabilities in source code. We discuss challenges and difficulties of this setting as well as recent results of our approach for "vulnerability extrapolation".

5.7 Using Machine Learning in Digital Forensics

Felix C. Freiling (Universität Erlangen-Nürnberg, DE)

Joint work of Freiling, Felix C.; Dewald, Andreas; Kälber, Sven; Rieck, Konrad; Ziehe, Andreas

We report on our first experiences of using machine learning for digital event reconstruction in the context of digital forensics. The idea is to learn typical patterns of system/user activities in file system metadata and use this information to infer or classify given file system images with respect to such activities.

5.8 Kernel Mode API Spectroscopy

Viviane Zwanger (Universität Erlangen-Nürnberg, DE)

Joint work of Zwanger, Viviane; Freiling, Felix
Main reference V. Zwanger, F.C. Freiling, "Kernel Mode API Spectroscopy for Incident Response and Digital Forensics," PPREV 2013 (ACM SIGPLAN Program Protection and Reverse Engineering Workshop), Rome.

The new generation of rootkits is heading towards well-designed and carefully crafted attacks to industrial and safety-critical systems with the objective of collecting specific data from a certain target. This creates a need for detecting and analyzing this kind of malicious behavior. We present a simple and surprisingly effective technique which we call "API spectroscopy" to quickly assess the nature of binary code in memory in an automated way. We apply API spectroscopy to the problem of analyzing Windows kernel drivers and kernel mode rootkits. Our method scans the binary code for calls to kernel mode API functions and outputs a histogram of these calls with respect to different semantic classes. We call the result the API spectrum of the driver and the method API spectroscopy. When API calls are grouped into functional classes, an API spectrum can give a compact insight into the possible functionality of an unknown piece of code and therefore is useful in IT incident response and

digital forensics. Examples for such semantic classes are "networking","filesystem access", "process management" or "DMA/Port I/O". We present the design and implementation of an API spectroscope for the Windows operating system. We tested different legitimate drivers from the categories "networking", "filesystem" and "Hardware" as well as different malicious kernel mode rootkits found in the wild: Duqu, the German Police rootkit software, a facebook-account stealer rootkit and a newly discovered yet unknown rootkit. The resulting API spectra were found to reflect the behavior of the analyzed driver quite well and might be used as a fingerprint system to detect new rootkits.

5.9 Training Data Sanitization in Adversarial Learning: Open Issues

Patrick Pak Kei Chan (South China University of Technology, CN)

The goal of pattern classification is to generalize the knowledge from the training data to the unseen samples. Unfortunately, in many real applications, the training set is influenced by an attacker. The influenced dataset misleads the learner and downgrades its generalization ability. One method to solve this problem is data sanitization, which identifies and eliminates the attack data from the training set before learning. In this talk, the noisy and attack data sanitization concepts are compared. The preliminary idea of sanitizing the training data in adversarial learning is presented.

6 Working Groups

6.1 Machine Learning for Computer Security

Battista Biggio (Università di Cagliari – Cagliari, Italy)
Nedim Šrndić (Universität Tübingen – Tübingen, Germany)

This workgroup explored the current role of machine learning in security research and the success and failures in using learning methods within security applications. This group identified open issues and research priorities including the need for transparency, interpretability and trust for secure learning approaches, the need for preventive measures and the potential for using learning in penetration testing, and the need for scalable procedures. They also identified future applications for secure learning including detecting advanced persistent threats, dynamic authentication, autonomous monitoring, and crime prediction.

6.2 Secure Learning: Theory and Methods

Daniel Lowd (University of Oregon – Eugene, OR, USA)
Rachel Greenstadt (Drexel University – Philadelphia, PA, USA)

The secure learning workgroup confronted the challenges that face learning researchers, who design learning algorithms for adversarial environments. They discussed how security can be formulated as an objective for designing learning procedures; the need for security metrics; the need for developing security-driven benchmarks and adversarial data simulations for evaluating learning approaches; and general techniques, constraints and challenges for developing secure learning technologies. Finally this group identified a set of key open questions including what should be achieved by secure learning, how can we know the adversary, and what is the appropriate role of secure learning within a secure system?

6.3 Future Applications of Secure Learning

Nathan Ratliff (Google – Pittsburgh, PA, USA)
Alvaro Cárdenas (Fujitsu Labs of America Inc. – Sunnyvale, CA, USA)
Fabio Roli (Università di Cagliari – Cagliari, Italy)

The final workgroup addressed the task of identifying future applications, which could benefit from these secure learning technologies. This group examined domains including intrusion detection, malware analysis, spam filtering, online advertisement, social media spam, plagiarism detection / authorship identification, captcha cracking, face detection, copyright enforcement, and sentiment analysis. The group also compiled a list of major research priorities including the need to address poisoning attacks and the need for benchmarks and penetration testing. They also highlighted privacy, non-stationarity of data, and the lack of ground truth as the major hindrances to the production of adequate benchmark datasets for secure learning.

7 Overview of Open Problems

In the "Machine Learning Methods for Computer Security" Perspectives Workshop, attendees identified the following general issues that need to be addressed by the community:

The need for learning approaches to provide a greater degree of trust for seamless integration of these approaches in security applications

There is generally an apprehension within the security community toward black-box learning procedures that can be addressed by providing greater transparency, interpretability and trust. There is, further, a need for preventive and corrective learning procedures and for using machine learning in an offensive role such as using learning to perform penetration testing to better validate existing techniques.

Design problems for incorporating secure learning technologies into security applications

There is also a general question of identifying the appropriate role of secure learning in secure systems and developing a methodology for designing such systems. Among the challenges for this task are finding scalable learning procedures that can meet the security objectives and incorporating human operators to help provide security safeguards.

The need for a more methodical approach to secure learning

Currently, the notion of secure learning largely remains ill-defined and domain-specific and lacks proper evaluation. There is a need to incorporate a realistic model of the adversaries, to formalize secure learning, and to identify which applications can benefit from these technologies. Stronger connections between secure and private learning would be desirable and perhaps result in unifying these currently disparate fields.

Construction of benchmarks and case studies for secure learning

Several groups identified a need for and current lack of adequate benchmarks and case-studies for evaluating secure learning. Unfortunately, real-world security data remains scarce for reasons including the concern of many security-sensitive systems for safeguarding their users' privacy, the ever-evolving nature of the data in these typically non-stationary domains, and the lack of ground truth due to the fact that many attacks may never be revealed or discovered.

Acknowledgments

We thank Dr. Roswitha Bardohl, Susanne Bach-Bernhard, Dr. Marc Herbstritt, and Jutka Gasiorowski for their help in organizing this workshop and preparing this report.

Participants

- Battista Biggio
Università di Cagliari, IT
- Christian Bockermann
TU Dortmund, DE
- Michael Brückner
SoundCloud Ltd., DE
- Alvaro Cárdenas Mora
Fujitsu Labs of America Inc. – Sunnyvale, US
- Christos Dimitrakakis
EPFL – Lausanne, CH
- Felix C. Freiling
Univ. Erlangen-Nürnberg, DE
- Giorgio Fumera
Università di Cagliari, IT
- Giorgio Giacinto
Università di Cagliari, IT
- Rachel Greenstadt
Drexel Univ. – Philadelphia, US
- Anthony D. Joseph
University of California – Berkeley, US
- Robert Krawczyk
BSI – Bonn, DE
- Pavel Laskov
Universität Tübingen, DE
- Richard P. Lippmann
MIT – Lexington, US)
- Daniel Lowd
University of Oregon, US)
- Aikaterini Mitrokotsa
EPFL – Lausanne, CH
- Sasa Mrdovic
University of Sarajevo, SEU
- Blaine Nelson
Universität Tübingen, DE
- Patrick Pak Kei Chan
South China University of Technology, CN
- Massimiliano Raciti
Linköping University, SE
- Nathan Ratliff
Google – Pittsburgh, US
- Konrad Rieck
Universität Göttingen, DE
- Fabio Roli
Università di Cagliari, IT
- Benjamin I. P. Rubinstein
Microsoft – Mountain View, US
- Tobias Scheffer
Universität Potsdam, DE
- Galina Schwartz
University of California – Berkeley, US
- Nedim Srndic
Universität Tübingen, DE
- Radu State
University of Luxembourg, LU
- Doug Tygar
University of California - Berkeley, US
- Viviane Zwanger
Univ. Erlangen-Nürnberg, DE

Report from Dagstuhl Seminar 12372

Biological Data Visualization

Edited by

Carsten Görg[1], Lawrence Hunter[2], Jessie Kennedy[3], Seán O'Donoghue[4], and Jarke J. van Wijk[5]

1 **University of Colorado, US,** carsten.goerg@ucdenver.edu
2 **University of Colorado, US,** larry.hunter@ucdenver.edu
3 **Napier University – Edinburgh, GB,** j.kennedy@napier.ac.uk
4 **CSIRO and the Garvan Institute of Medical Research, AU,** sean@odonoghuelab.org
5 **Eindhoven University of Technology, NL,** vanwijk@win.tue.nl

Abstract

The topic of visualizing biological data has recently seen growing interest. Visualization approaches can help researchers understand and analyze today's large and complex biological datasets. The aim of this seminar was to bring together biologists, bioinformaticians, and computer scientists to survey the current state of tools for visualizing biological data and to define a research agenda for developing the next generation of tools. During the seminar, the participants formed working groups on nine different topics, reflected on the ongoing research in those areas, and discussed how to address key challenges; six talks complemented the work in the breakout groups. This report documents the program and the outcome of Dagstuhl Seminar 12372 "Biological Data Visualization".

Seminar 09.–14. September, 2012 – www.dagstuhl.de/12372
1998 ACM Subject Classification H.5 Information Interfaces and Presentation, I.3 Computer Graphics, J.3 Biology and genetics
Keywords and phrases Information visualization, data visualization, biology, bioinformatics, user interfaces, visual analytics
Digital Object Identifier 10.4230/DagRep.2.9.131

1 Executive Summary

Carsten Görg
Lawrence Hunter
Jessie Kennedy
Seán O'Donoghue
Jarke J. van Wijk

Introduction and Motivation

Biology is rapidly evolving into a 'big data' science, and as a consequence there is an urgent and growing need to improve the methods and tools used for gaining insight and understanding from biological data. Over the last two decades, the emerging fields of computational biology and bioinformatics have led to significant advances primarily in automated data analysis. Today, however, biologists increasingly deal with large, complex datasets (e.g., 'omics' data) where it is not known in advance what they are looking for and thus, automated analyses alone

cannot solve their problems. Interactive visualizations that can facilitate exploratory data analysis and support biologists in creating new hypotheses lend themselves to complement automated analyses. Bioinformaticians already have built a variety of tools for visualizing different types of biological data and those tools are widely used in the community. So far, most bio-related visualization research has been conducted by people outside of the visualization community, people who have learned about visualization but are often not aware of research in the visualization community. Consequently, the current tools do not embody the latest advancements in design, usability, visualization principles, and evaluation.

One main goal of this first Dagstuhl Seminar on Biological Data Visualization was to bring together the users (biologists), current visualization tool builders (bioinformaticians), and visualization researchers to survey the state-of-the-art of the current tools and define a research agenda for systematically developing the next generation of tools for visualizing biological data. Only a close collaboration of the researchers from all three communities can create the synergies necessary to address the challenges in analyzing and visualizing large and complex biological datasets.

Topics discussed during the seminar included:

- Challenges in visualizing biological data. Biological data is very heterogeneous. It contains spatial data, graphs, tabular data, and textual data. Challenges are wide spread: open-ended data quantity, open-ended exploratory tasks, long-term analyses, rich analytics, heterogeneous data, usability and evaluation of tools.
- Design and visualization principles, research in human-centered design, usability, and evaluation of interactive data-analysis tools.
- Creating a common research agenda and a common understanding of the problem field of biological data visualization.
- Integration of multiple visualizations for different data types and tasks into one tool to support more complex analysis scenarios.
- Designing an infrastructure for next generation visualization tools.
- Establishing collaborations between computer scientists and biologists.

Participants and Program

41 researchers from 9 countries participated in this seminar. Many participants came from the US and from Germany, others came from Canada, Australia, and a number of other European countries (see Figure 1). There was a good mix of researchers from the visualization, bioinformatics, and biology communities. About a third of the participants attended their first seminar at Dagstuhl.

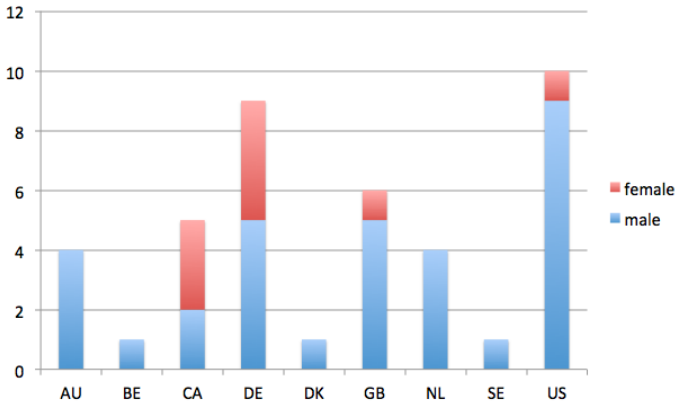


Figure 1 Participant statistics of the seminar.

Monday	Tuesday	Wednesday	Thursday	Friday
Introduction Personal Ads	Talk Reporting Session	Talk Reporting Session	Talk Breakout Groups	Reporting Session
Discussion: Topics for Breakout Groups	Breakout Groups	Discussion: Topics for new Breakout Groups	Breakout Groups	Discussion: BioVis Community
Talk Breakout Groups	Talk Breakout Groups	Excursion to Trier	Talk Breakout Groups	
Breakout Groups	Breakout Groups		Breakout Groups	

Table 1 Final schedule of the seminar. The breakout groups on Monday/Tuesday discussed topics on Ontologies in Biological Data Visualization, Comparative Analysis of Heterogeneous Networks, Sequence Data Visualization, and Bridging Structural & Systems Biology; the breakout groups on Wednesday/Thursday discussed topics on Uncertainty Visualization, Infrastructure, Multiscale Visualization, Effective Visualization Design, and Evaluation.

Table 1 provides an overview of the final seminar schedule. The program was designed to facilitate in-depth discussions in small working groups. To get to know each other—the seminar brought together researchers from different communities—participants introduced themselves and their research interests with a 'personal ad' in the Monday morning session. This was a great way to set the tone for informal and engaging discussions during the seminar.

Previous to the seminar, the organizers collected interesting ideas and suggestions from the participants for possible topics for working groups. To allow participants to work on different topics and with different people, the topics and groups changed halfway through the seminar. On Monday morning and Wednesday morning all participants discussed and refined the suggested topics and formed groups according to their interests. The groups (four on Monday/Tuesday and five on Wednesday/Thursday) worked in parallel on their topics and reported regularly on their progress. The work in the breakout groups was complemented by a discussion on the BioVis Community on Friday and a number of talks given throughout the seminar:

- *Seán I. O'Donoghue*: BioVis Introduction: A Practitioner's Viewpoint
- *Daniel Evanko*: Visualization on nature.com
- *Matt Ward*: Biovisualization Education: What Should Students Know?
- *Arthur J. Olson*: The Promise and Challenge of Tangible Molecular Interfaces
- *Martin Krzywinski*: visualization – communicating, clearly
- *Bang Wong*: Concepts gleaned from disparate communities

These talks, presented to all participants in the morning sessions and after the lunch breaks, intentionally touched on broad and high-level topics to make them more interesting to the diverse audience in the seminar. The abstracts of the talks are presented in Section 3.

Discussion and Outcome

Some of the working groups followed a classical design process [6, 8] to structure their collaborative work. They split their discussions into a *problem phase* and a *solution phase*. Both phases featured divergent and convergent stages: *discover* and *define* for the problem phase and *develop* and *deliver* for the solution phase. Francis Rowland, a seminar participant with expertise in user experience design, facilitated these discussions.

Figure 2 shows some artifacts produced by the Ontologies in Biological Data Visualization working group that followed this design process. The *Four C's* approach (left) is an example for the discover and design stages. The group broke down their topic into four aspects: *Components* (parts), *Characters* (people involved), *Challenges*, and *Characteristics* (features and behavior). The Four C's approach helped the group to provide a holistic view on the design problem and to better define the topic. The *Draw the Box* approach (right) is an example for the develop and deliver stages. Members of the group collaboratively imagined an end product of their work that would be sold in a box on a shelf and designed its package. This approach helped the group members to gather ideas, visualize the outcome, and focus on the most important features of the product.

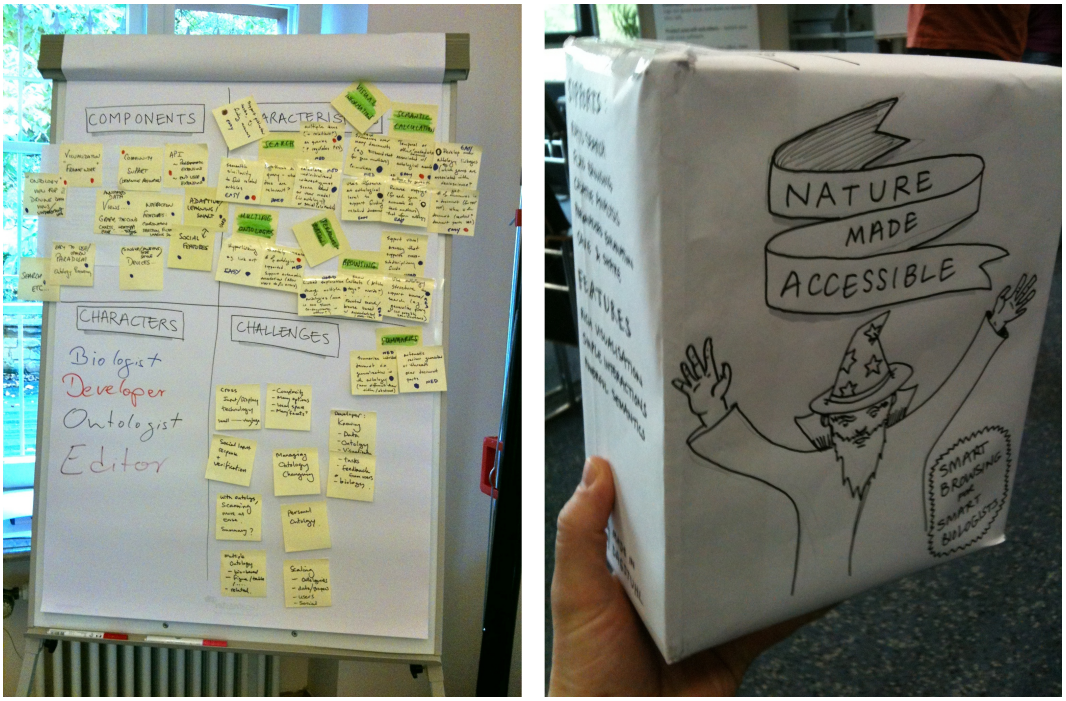


Figure 2 Examples of design processes: the *Four C's* approach (left) and the *Draw the Box* approach (right).

The diverse outcomes from the nine working groups are summarized below. The detailed reports are presented in Section 4.

Comparative Analysis of Heterogeneous Networks: The analysis of the transcriptome produces a large number of putatively disrupted transcripts, and prioritizing which disruptions are most likely to be meaningful (causal or diagnostic) is a time-consuming process. To guide their interpretation researchers create heterogeneous networks by integrating information from a wide variety of annotation databases. The working group investigated how the analysis of the transcriptome can be facilitated by interactive visualizations of transcriptome assemblies and proposed a method to infer the functional consequence of a transcript's disruption based on the local structure of the annotation networks. A tight coupling of network analysis algorithms and interactive visualizations, specifically designed to support these analysis tasks, could accelerate identification of important transcript alterations.

Sequence Data Visualization: Genome-associated data is growing at a fast rate and genome browsers are still the tool of choice for integrating and analyzing different types

of data in one single representation. The working group analyzed the different challenges of visualizing genome-associated data and separated them into two different dimensions: problems associated with rearrangements of the genomic coordinates and problems with the abundance of data at each genomic position. To address these problems, the group discussed and developed a number of possible solutions, including the development of a reference-free gene-centric approach, compressing tracks by aggregation or summarization, and using meta-data or data itself as a novel way for selecting tracks. These approaches can lay the foundation for the development of new visualization tools.

Bridging Structural & Systems Biology via DataVis: There exist several gaps between the field of structural biology, which has yielded detailed insight into the molecular machines of life, and the field of systems biology, which has evolved more recently in the wake of the genomics revolution, but separately from the advances of the more structural view of biology. The integration of both fields and their visualization tools could create new tool sets to enhance the exploration and understanding of biological systems. The working group analyzed and described the existing gaps and proposed seven strategies to facilitate collaboration and professional advancement in structural biology, systems biology, and data visualization.

Ontologies in Biological Data Visualization: Ontologies are graph-based knowledge representations in which nodes represent concepts and edges represent relationships between concepts. They are widely used in biology and biomedical research, for the most part as computational models, in computational analyses, and for text mining approaches. The working group examined the potential impact of ontologies on biological data visualization. The group identified challenges and opportunities from the perspectives of three different stakeholders: ontologists (who create and maintain ontologies), data curators (who use ontologies for annotation purposes), and data analysts (who use ontologies through applications to analyze experimental data). Identified challenges include the dynamic nature of ontologies, scalability, how to utilize the complex set of relationships expressed in ontologies, and how to make ontologies more useful for data analysis. Identified research opportunities include the visualization of ontologies themselves, automated generation of visualization using ontologies, and the visualization of ontological context to support search. The group submitted a Viewpoints article on Ontologies in Biological Data Visualization to the IEEE Computer Graphics & Applications journal.

A Framework for Effective Visualization Design: Visualizations are not only an important aspect of how scientists make sense of their data, but also how they communicate their findings. The techniques and guidelines that govern how to design effective visualizations, however, can be quite different whether the goal is to explore or to explain. Unfortunately, scientists are often not aware of the spectrum of considerations when creating visualizations. To help clarify this problem, the working group has developed a framework to reason about the spectrum and considerations to help scientists better match their visualization goals with appropriate design considerations.

Uncertainty Visualization: Uncertainty is common in all areas of science, and it poses a difficult problem for visualization research. Visualization of uncertainty has received much attention in the areas of scientific visualization and geographic visualization; however, it appears much less common in information visualization and in biological data visualization. The working group analyzed and described the sources of uncertainty and types of uncertainty specific to biology. Uncertainty visualization in networks was identified as an open issue, including uncertainty in the network topology and uncertainty in attributes on nodes, edges, and their interdependencies. The group started a survey of the literature on uncertainty

visualization for biological data and proposed to construct a taxonomy of uncertainty visualization approaches, and investigate how they could be employed in the context of a collection of biological problems.

Evaluation: The working group identified two central problems with respect to the evaluation of tools for visualizing biological data: (1) How to motivate biologists to participate in evaluations? and (2) How to evaluate the tools? The answer to the first question was (simply) that biologists have to benefit from the evaluation to be motivated to participate, e.g. they might get a tool they can use to solve their problems. The second question was more complex and the working group discerned a number of dimensions, centered around what, why, when, where, and how. The discussion of these dimensions lead to the insight that there is a strong difference between approaches taken by designers working at a bio-institute and approaches taken by infovis researchers. Both approaches have merit, the challenge is to close the gap and combine them.

Multiscale Visualization: Biology involves data and models at a wide range of scales and researchers routinely examine phenomena and explore data at multiple scales. Visual representations of multi-scale datasets are powerful tools that can support data analysis and exploration, however, visualizing multi-scale datasets is challenging and not many approaches exist. The working group identified four common dimensions of biological multi-scale datasets: 3D space, time, data complexity (modality), and data volume (size). The group produced a short video to introduce each dimension independently in order to provide a quick and understandable view on the nature of the different scales and how they apply to biological data and exploration. Additionally, the group discussed in more detail a number of biological multi-scale data and models that can be visualized across multiple dimensions and introduced case studies to highlight issues like navigation, interaction, and human-computer interfaces. Carsten Görg presented a talk on the results from this working group at the 2012 Rocky Mountain Bioinformatics Conference.

Infrastructure: The working group discussed needs from both a technical and community standpoint regarding the challenges involved in the analysis of biomedical data and mechanisms to facilitate interactions between visualization communities in computer science and biology. Eight key criteria were identified: interoperability, reusability, compatibility, references & benchmarks, middleware, vertical integration, scalability, and sustainability. The group developed a model for a community-maintained, biological visualization resource that would enable biological questions, task descriptions, sample datasets and existing tools for the problems to be disseminated to the computational visualization and biological research communities. Additionally, the group developed a detailed use-case based on the data and analysis pipelines of the cancer genome atlas that will allow technical aspects of the eight key criteria to be explored and practical solutions proposed.

Finally, based on feedback from the participants (from the seminar questionnaire as well as from personal communication with the organizers) another important outcome of the seminar was to establish collaborations between computer scientists and biologists. The academic cultures in biology and computer science, including publication models, are quite different. In addition, biologists have a different mindset than computer scientists: biologists often work in a detail-oriented manner whereas computer scientists often seek to generalize. Understanding each other's culture is important for successful collaborations and the Dagstuhl seminar provided a unique setting to meet enthusiastic people from different communities, have long group discussions with a focus on problem solving, and form synergies with researchers that have a different outlook and expertise.

2 Table of Contents

3 Overview of Talks

3.1 BioVis Introduction: A Practitioner's Viewpoint

Seán I. O'Donoghue (CSIRO and the Garvan Institute of Medical Research, AU)

Experimental methods in biological research are delivering data of rapidly increasing volume and complexity. However, many current methods and tools used to visualize and analyse these data are inadequate, and urgent improvements are needed if life scientists are to gain insight from this data deluge, rather than being overwhelmed. I will discuss a recent switch in focus away from algorithmic bioinformatics towards data visualization and usability principles, illustrating how such a focus can have significant impact, illustrating these points with examples from work on macromolecular structures, systems biology, and literature mining. I will also discuss a recent, international community initiative that brings visualization experts together with computational biologists, bioinformatics, graphic designers, animators, and medical illustrators, and aims to raise the global standard of bioinformatics software (http://vizbi.org/).

3.2 Visualization on nature.com

Daniel Evanko (Nature Publishing Group, US)

Nature very recently published results of the Encyclopedia of DNA Elements (ENCODE) project. To aid the discoverability of information in these manuscripts Nature Publishing Group developed the ENCODE Explorer and threaded presentations of the results to allow targeted reading of single selected topics through all 36 manuscripts of the project. We also created Javascript-based interactive figures with the intention of further developing and reusing these visualizations elsewhere. As a further aid to information discoverability, technical editors are beginning to annotate all gene, protein and chemical entities in original research papers published in a limited number of Nature research journals. We hope to make this information accessible through APIs.

3.3 Biovisualization Education: What Should Students Know?

Matt Ward (Worcester Polytechnic Institute, US)

In recent years, the level of activity in the area of visualization of biological data has greatly increased, both in terms of users and developers. An important question is what sort of training should be provided for students in this area? Can we use existing courses in biology and data/information visualization, or do we need one or more courses that fuse these two distinct fields? In this talk I describe my experiences in designing and delivering a course

on biovisualization to upper level undergraduate and graduate students majoring in either computer science or bioinformatics and computational biology. The project oriented course covers both basic principles of information visualization as well as the data models typically found in biology – sequences, networks, tabular data, and spatial structures. For each data type I describe the common analysis tasks performed on the data as well as a variety of visual mappings that can be applied. I also describe standard rules for effective visualization design and common methods for evaluating the resulting visualizations. I summarize my observations on the best and weakest aspects of the course and welcome feedback from the seminar attendees on ways to improve the course.

3.4 The Promise and Challenge of Tangible Molecular Interfaces

Arthur J. Olson (The Scripps Research Institute – La Jolla, US)

Structural molecular biology is a key science in connecting the worlds of physics and chemistry to biology. It is a discipline that focuses on three and four-dimensional relationships of complex shapes and functions. As such, it has been a fertile proving ground for novel technologies that can enhance interaction and visualization of such systems for the purposes of exploration, understanding and communication.

Physical models have been used for centuries to aid in the process of modeling and visualization in many areas of science. In the latter part of the last century computer graphics largely superseded physical models for these purposes. This advance in technology was accompanied by a loss of the perceptual richness inherent in the human interaction with real physical objects. The tactile and proprioceptive senses provide key cues to our ability to understand 3 dimensional form and to perform physical manipulations, but are now currently under-utilized in fields such as molecular biology.

We have been developing new ways to represent, visualize and interact with the molecular structures that make up the machinery of life. We are adapting two emerging computer technologies, *solid printing* and augmented reality, to create a natural and intuitive way to manipulate, explore and learn from molecular models. We create tangible models utilizing computer autofabrication. Each model can be custom made, with an ease similar to that of printing an image on a piece of paper. Specific model assembly kits can be made with this technology to create *molecular Legos* that go well beyond the chemical models of the nineteenth and twentieth centuries. Augmented reality is used to combine computer-generated information with the physical models in the same perceptual space. By real-time video tracking of the models as they are manipulated we can superimpose text and graphics onto the models to enhance the information content and drive interactive computation.

These models and tangible interfaces have been used in both research and educational settings. The talk will include a live demonstration of the models and interactive use in an augmented reality setting.

3.5 visualization – communicating, clearly

Martin Krzywinski (BC Cancer Research Centre, CA))

We should think about visualization not only in terms of effective data encodings, but also in terms of design. We use visualizations to communicate patterns and concepts and are more effective if we incorporate design principles in our figures. Clutter and redundancy can muddle a figure – two pitfalls into which many figures fall. Using examples of redesigned figures, I will motivate how mitigating these two issues can improve visual communication. I will also work through a Nature figure redesign in detail to demonstrate the process and how you can apply it in your workflow.

3.6 Concepts gleaned from disparate communities

Bang Wong (Broad Institute of MIT & Harvard – Cambridge, US)

4 Working Groups

4.1 Comparative Analysis of Heterogeneous Networks

Andreas Kerren, Corinna Vehlow, Jessie Kennedy, Karsten Klein, Kasper Dinkla, Michel Westenberg, Miriah Meyer, Mark Ragan, Martin Graham, Martin Krzywinski, and Tom Freeman

Transcriptome sequencing of a large cohort is now a routine method of interrogating the profile of expressed gene products of many individuals. Analysis of the transcriptome produces a large number of putatively disrupted transcripts, and prioritizing which disruptions are most likely to be meaningful (causal or diagnostic) is a time-consuming process. Researchers integrate information from a wide variety of annotation databases, many of which are interaction or pathway networks, to guide their interpretation of how a disrupted transcript might affect the functioning of a cell.

We investigated how this analysis can be facilitated by interactive visualizations of transcriptome assemblies in the context of these networks. The goal of our proposed method is to infer the functional consequence of a transcript's disruption based on local structure of the annotation networks (Figure 3). For example, the investigator may have identified functional motifs in the network that are relevant to their hypotheses, and conclude that any disrupted transcripts found in these motifs are likely to be important.

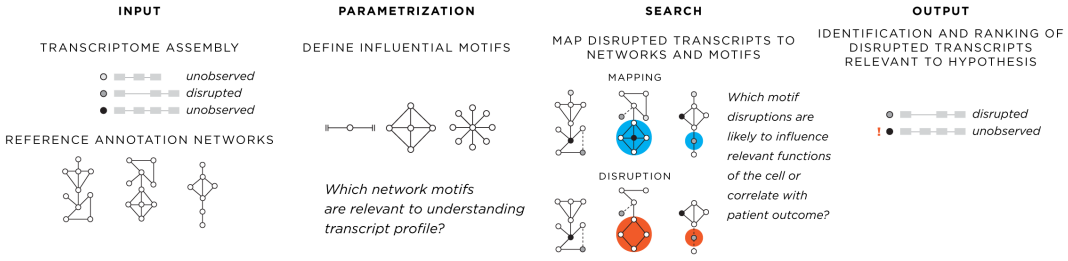


Figure 3 Transcript assemblies from the sequencing of a cancer genome produce indication of disrupted or unobserved transcripts. The existence of reference annotation networks onto which these transcripts can be mapped provides a method of assessing the functional implications of the disruption.

We claim that a tight coupling of network analysis algorithms and interactive visualizations, specifically designed to support these analysis tasks, would accelerate identification of important transcript alterations. A software system that realizes such a coupling could streamline the process of identifying influential network motifs, determining whether disrupted transcripts fall within these motifs, and support the process of deriving a priority rank based on the results of this search. By using a system of linked views, each showing one of the reference networks, the system could provide the researcher a means of mapping transcript disruptions onto the networks. The views would show constrained locales of each motif to decrease the information burden — the reference networks are typically very large (10,000+ nodes) — and preserve the users' mental map. This concept supports the analysis of disruptions in the context of different reference networks at a time and therefore helps users to assess the impact of the disruption.

Our plan is to formalize this analysis method and implement a software system to realize it in practice.

4.2 Sequence Data Visualization

Jan Aerts, Jean-Fred Fontaine, Michael Lappe, Raghu Machiraju, Cydney Nielsen, Andrea Schafferhans, Svenja Simon, Matt Ward, and Jarke J. van Wijk

Introduction

Genome-associated data is growing at a fast rate. Since the advent of large genome sequencing efforts such as the Human Genome Project, the genome browser (e.g. UCSC genome browser, Ensembl) has been the tool bringing all types of data together into a single representation. This report serves to analyze the shortcomings of current genome browsers and to suggest options for solving the problems.We identified two main use cases in using genome browsers: hypothesis verification and hypothesis generation. In the first use case (*hypothesis verification*) users want to verify a hypothesis (e.g. the involvement of a certain gene in the development of cancer) by checking whether experimental data can support this hypothesis. Current genome browsers allow uploading custom experimental data in order to analyze it in context of other data. In the second use case (*hypothesis generation*), users often have access to much experimental data that needs to be interpreted and are looking for significant signals that fall into regions where there is evidence of functional relevance.Although generic genome browsers have clearly proven their worth, they are now starting to show clear shortcomings. These can be separated into two dimensions. First, using a *fixed genome coordinate system* works as long as one is only interested in the reference sequence and features that have fixed and clear positions on that reference sequence. When, however, considering structural genomic variations (i.e. duplications, deletions, inversions, and translocations) the paradigm of annotation vis-a-vis a fixed reference starts to break down. Second, the current concept of displaying features in different *tracks* becomes cumbersome with the immense growth of annotation data. The increasing number of annotation tracks to be selected/deselected for display makes keeping an overview of the available data nearly impossible.In addition to the issues when considering structural variation or large track lists, the integration of uncertainty in feature visualization is still lacking. This uncertainty exists at two different forms: statistical uncertainty and positional uncertainty. *Statistical uncertainty* reflects the confidence that one has towards the existence or correctness of that feature or not. Therefore, it is important to not only view summarizing annotations, but also to be able to investigate the underlying evidence. The representation of statistical uncertainty was the topic of a separate Dagstuhl breakout group, and therefore not further considered within the current group. *Positional uncertainty* considers the resolution of feature annotation. The boundaries (breakpoints) of deletions, for example, can often not be identified exactly, but can be known to lie within a certain range. At present this type of information is not displayed in the generic genome browsers.Below we examine the two dimensions to the problem of displaying genomic information, namely the reliance on a single genomic coordinate system and the large number of feature tracks.

Visualizing genomic structural variation

A structural variant consists of a DNA sequence, typically >1 kilobase, that deviates from a reference sequence in content, order and/or orientation. A distinction can be made between balanced variations (i.e. inversions and translocation) that do not change the total genome content, and unbalanced variations (i.e. deletions and duplications) that do result in a change of the total genome content (see Figure 4). The latter therefore are also known as copy number variations or CNVs. Although detection of structural variations has long been possible using e.g. fluorescent *in situ hybridization* (FISH) or array comparative genome hybridization (aCGH), they are now routinely detected using next-generation sequencing (NGS) technology. After paired-end sequencing of one or more samples and mapping these reads to the reference genome, patterns in read depth as well as aberrant distance between and/or orientation of paired-end sequences can indicate structural variations.

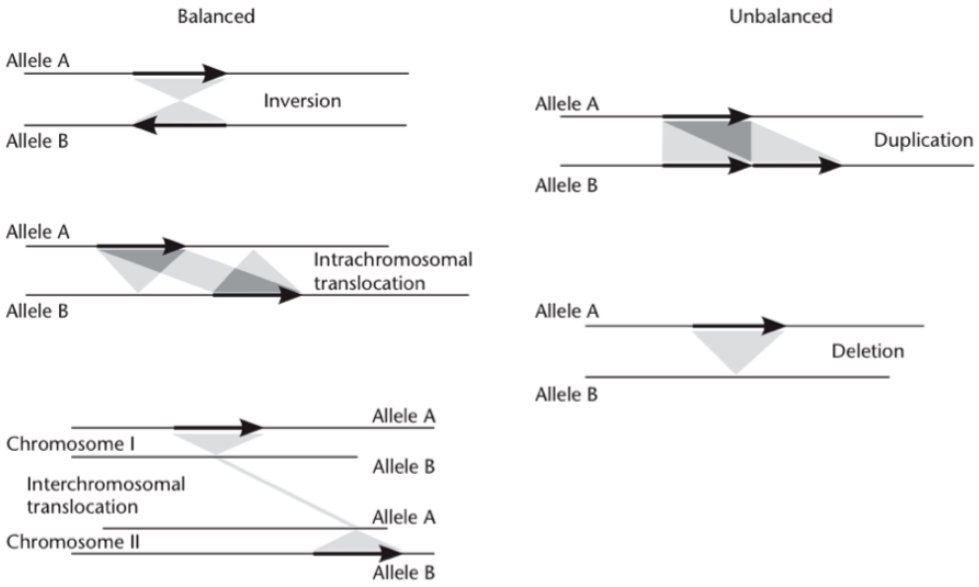


Figure 4 Types of structural genomic variation (taken from [1]).

Problem Statement

The issue with representing structural variation using generic genome browsers is two-fold. First, the data to be represented (both underlying read mapping data and resulting variations) often involves features that are linked at two *different loci in the genome*. Read pairs, for example, consist of two reads that do belong together but can be mapped to very distant regions in the genome. Duplications and translocations also inherently constituted of two elements: the locus that acts as the source for the duplication/translocation, and the locus that acts as the target. Some efforts have been made to resolve this issue, e.g. by providing a split-pane view on the data (Integrative Genome Viewer; Robinson et al, 2011). There is however a second and more profound issue of reliance on a *single reference genome*. Any variation between a sample and the reference can only be displayed as a feature in a track of the genome browser (see Figure 5). As a result, two different samples cannot be compared directly to each other, but can only be compared by how each differs from the reference. In

addition, the reference-genome based representation does not reflect the *in vivo* configuration of the sample chromosome. For example, a region of the reference genome that is deleted in a sample can be highlighted with colored bars in a genome browser (see red bars in Figure 5). But this requires that the user build a mental-model of which genomic regions are now adjacent as a result of the deletion in the sample rather than being able to observe the new junction directly. This is fairly straightforward for simple variants, but quickly becomes challenging for more complex structural changes.

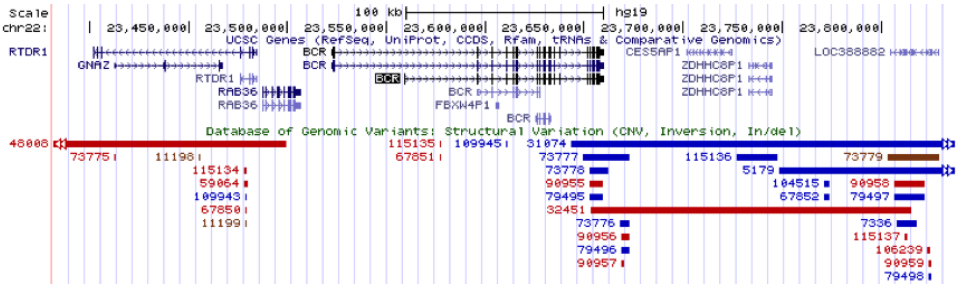


Figure 5 Structural variations annotated around the BCR gene on chromosome 22 (picture taken from the UCSC genome browser). Duplications are in blue, deletions in red, inversions (not shown here) in purple.

Our Approach

Some inroads have been made into representing features with more than one position (e.g. read pairs, or duplications and translocations). Visualization approaches and software tools have been reviewed previously ([13, 12]). Most notable is the use of the Circos viewer ([11]). This circular visualization maps the different chromosomes onto a circle and links related loci through the use of bezier curves (see Figure 6). Another approach is the dot plot in which the x and y axes correspond to the two sequences being compared, and points indicate sequence identity. Diagonal lines indicate corresponding sequence segments and the horizontal offset highlights reordering.

Both circular and dot plot representations emphasize the positions of structural variants on the genomic coordinate. Since the start of the genome browsers, the main nugget of information to be displayed has always been positional. Researchers have become accustomed to this habit, and novel visualization concepts are necessary. One option is to consider the genome as a collection of functional elements rather than a linear scaffold and emphasize the biological consequences of the variants rather than their genomic arrangement. Indeed, the location of a functional element (i.e. a gene together with any *cis*-acting regulators) on a chromosome is irrelevant for most purposes. A reference-free gene-centric approach will therefore be developed where the emphasis is on the contents of the genome rather than on its linear structure. In addition, a genome can be represented as a collection of segments that can be rearranged between different individuals in a graph-like structure. We can draw inspiration from previous work in this area ([14, 7]). By combining these representations with a circularized linear layout (such as Circos), we believe that a researcher can build a comprehensive overview of the effects of a structural variation.

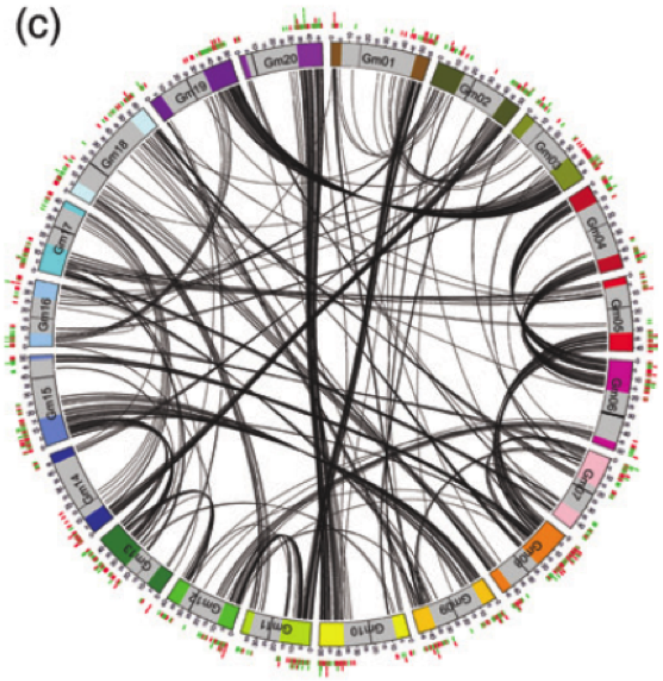


Figure 6 Example of Circos (taken from [17]).

Track compression

Many different annotations and measurements are associated with genomic positions. Since not all available data will fit on a screen, let alone be interpretable to a user, some compression has to take place. This can be achieved by selecting relevant data or by aggregating or summarizing different tracks. In this section, we analyze the chances and challenges of these approaches.

Data Description

The data associated with sequences can be separated into two main types: qualitative or enriched description of a region or quantitative data associated with positions. The region definition usually refers to the reference genome. In cases where one aims at comparing data stemming from different genomic sequences, mapping the data to a common reference scheme can already be a challenge (refer to previous section). The label usually refers to a function of the genomic region, e.g. "protein coding region" referring to the encoded protein, or a summary of quantitative data. In order to be informative to a user not familiar with the different data types, the label often needs to refer to several disparate pieces of information. Quantitative data usually detail the value of measurements per residue position. Different tracks can refer to very different types of measurements, (e.g. expression data) or type of sequenced sample (e.g. tissue or disease group). This can be a challenge to aggregating quantitative data, because accumulating or averaging may not be appropriate.

Track Selection

One approach to managing the multitude of data is to select that subset of data that is relevant to the subject under investigation. An ideal track selection tool would help the user select tracks relevant to their research, which can then be displayed in a genome browser. However, this selection is not trivial:

- The data sets are usually characterized by a concise name that is helpful for the respective domain experts. But for exploration it is hard to find out the content that might be relevant to a specific question. Even if one has found a signal in a region of interest, it is not obvious what that signal means. Reading through all linked data descriptions is very tedious.
- There might be significant overlap and redundancy in the data. Therefore, adding more tracks does not always add more information. In these cases aggregation might be possible. On the other hand information might be discovered in measurements the user is not aware of.

We identified two distinct approaches of selecting relevant tracks: using meta-data or using the data itself:

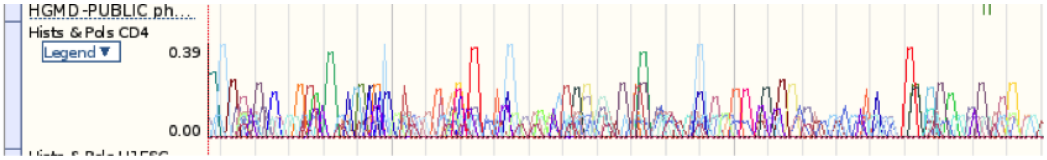
- The different tracks are currently characterized by names and descriptions. Various genome browsers use hierarchical organization and categorization to enable searching relevant data. However, we believe a more sophisticated use of controlled vocabulary or ontologies together with search systems could make the data more accessible. This categorization should allow to group data by different types of approaches, e.g. sample characterization (population, disease association, tissue, cell cycle), types of experiment (expression level, epigenetic modification), or type of summary annotation. An important requirement for this meta-data annotation is to allow grouping together tracks that can be aggregated without mixing apples and oranges.
- The data itself contains information that can help in selecting relevant selections or aggregation. Especially if users have experimental results they want to analyze in the context of the genomic information, looking for other signals that are statistically associated (correlated) with that data can help to find relationships and lead to new interpretations. This can also help to sort the tracks by relevance. Another related approach is to focus on region(s) of interest. Here, a simple filter can help to highlight those tracks that contain more than noise in the specified region. The remaining regions could be analyzed statistically to identify those tracks that show similar signals. On the one hand, this can help to aggregate related data to provide a more concise view; on the other hand the statistical analysis might point to common mechanisms governing the data and lead to new biological insight.

A very different approach to the selection of relevant tracks is the information contained in the community of users of the data. Very similar to the community recommendations used on sales web sites (e.g. Amazon) or social networks (e.g. LinkedIn), the data browser could suggest relevant tracks based on the usage statistics. Users with a high overlap of interest could make each other aware of interesting new directions to explore.

Data Aggregation

Another solution to making the data more manageable is aggregating different tracks into one. For example, if the user is only interested in finding out whether some variations fall into coding regions, all annotations of mRNA-mappings could be aggregated. However, here are a number of problems to solve:

- The semantical problem is to decide which tracks contain comparable information that may be aggregated, e.g. experimental results of a similar type or annotations of a similar type, but from different sources. Here, a good meta-data description, as described in the track selection section, is needed. This meta-data description should enable aggregation by different criteria, e.g. organized by cell or disease type.
- The visualization problem is how to show as much information as possible without overloading the user and without obscuring data. For example, if different quantitative measurements are shown simultaneously, using color or symbols to indicate the different measurements, it becomes hard to discern the different lines. Here a representation of e.g. the mean, the standard deviation, and specific outliers might be more helpful to show the most interesting aspects of the data.



Summary and Outlook

In the Dagstuhl workshop, we have analyzed the different challenges of visualizing genome-associated data and separated them into different dimensions: problems associated with rearrangements of the genomic coordinates and problems with the abundance of data at each genomic position. The problems and approaches for finding solutions outlined in this report will now be taken up in the development of new visualization tools. Although the discussion focussed on issues related to genomic sequence data, similar problems also exist in the realm of protein sequences. Therefore, the new concepts for visualizing data associated with genome sequences will hopefully also help to provide better overviews of protein sequences.

4.3 Bridging Structural & Systems Biology via DataVis

Graham Johnson, Julian Heinrich, Torsten Möller, Seán O'Donoghue, Art Olson, James Procter, and Christian Stolte

Introduction

Over 50 years of structural biology has yielded detailed insight into the molecular machines of life – from the scale of atoms to organs; the significance of this work with has been recognized by many Nobel Prizes. In contrast, systems biology has evolved over the past 20 years in the wake of the genomics revolution, but separately from the advances of the more structural view of biology. Visualization techniques for both structural and systems biology have both evolved in response to the need to analyze experimental data; in contrast, more general data visualization (datavis) approaches have evolved from a variety of application areas, where biology did not play a major role.

Describing the Gaps between Scientific Disciplines

Systems biology has traditionally utilized data ranging from genomics, proteomics, metabalomics, etc. which attempt to characterized in a systematic way the flow of molecular interaction and information/control. Structural biology, on the other hand, seeks to characterize the physical nature of such interactions and information flow by characterizing spatial and temporal structures. Connecting these two views is a challenge that requires techniques that have been developed on both sides of the gap. Likewise a gap exists between builders of computational tools from the biological community, and those from the computer science community. Within the visualization community, these gaps exist along several dimensions. Firstly, integrating the different data modalities and algorithmic approaches that arise from the structural and systems biology has been a significant challenge. Secondly, the viewpoint of the biologist focuses mostly on the biological question, while that of the visualization specialist focuses on the complete user experience. In addition, there also exist significant cultural gaps between these communities. The biology community and the visualization community publish in different ways, meet at different conferences, and evaluate their work using different criteria. Biology meetings are overloaded with data and urgent, important and unsolved problems, and unmet requirements – as a result, they are segmented into tiny subfields. By contrast, computer science meetings are data- and problem-hungry. These gaps are significant but surmountable, and bridging them holds the promise of pushing biology to the next level. The purpose of this white paper is to propose some strategies and tactics that may help to build these bridges.

Bridging the Gaps

The contrasting metrics for performance in each discipline mean that models for research dissemination do not allow sufficiently rapid transfer of new problems and new visualization solutions between the two domains. This is critical, however, and as a key outcome of this discussion, we define the following recommend strategies to facilitate collaboration and professional advancement in structural biology, systems biology, and datavis.

Mentoring and exchange programs amongst biological visualization, structural biology, and systems biology research groups. The most direct route to enable ideas and approaches to cross fertilize our fields is to facilitate interdisciplinary training. Orthogonal integration, where data visualization students and researchers are temporarily embedded in biology groups, will enhance the exchange of state of the art principles and approaches, and familiarize all parties with the tools, technology and data architectures involved.

Interdisciplinary conferences and symposia. The VIZBI and BioVis meetings already incorporate a range of mechanisms to encourage productive engagement between these communities; this could be strengthen by modeling other interdisciplinary meetings, such as ISMB. In addition, these mechanisms could be encouraged in the larger, more mainstream meetings in each field.

Co-localized hackathons and tutorial workshops. Much of the fundamental software tools used in structural biology today were created by bringing together specialists in numerical computation, physicists, physical chemists and biochemists. Focused workshops would enable engineers, theoreticians and applied researchers to identify new problems and design and prototype solutions informed by the latest visualization research. Whilst virtual participation is eminently feasible for these events, a one or two week period where specialists are physically co-located would maximize productivity. Tutorial sessions and facilitation

would be essential in these events to allow specialists to quickly understand and begin to apply their own knowledge to the problems at hand.

Exploitation of prepublication data repositories. Biological data repositories now play a key role for international collaboration, and could provide a means for data visualization researchers to access data and analysis problem solutions which would allow new solutions to be developed and evaluated in parallel with ongoing biological research programs.

Critical assessment of methodology exercises. Both communities have well established challenges that enable researchers to devise new solutions for data analysis problems. The model originally devised by Moult et al. in 1991 (http://www.predictioncenter.org) applied for the assessment of biomolecular structure prediction approaches has lead to a number of initiatives assessing approaches for biological text mining (http://www.biocreative.org) and biological systems reverse engineering (http://www.the-dream-project.org). These enterprises are distinct from the challenges in the data visualization field such as VAST and the BioVis challenge, since they employ real biological data which is accepted for publication but not yet released.

Clear definition of datavis challenges. If structural and system biologists were to clearly define specific visualization challenges, this would help the datavis community, enabling it to focus on more relevant problems that are likely to be adopted and to help advance the life science. The 2010 Nature Methods special issue (Vol. 7 No 3) and the ongoing VIZBI conference series are useful steps in this direction. The discussion group highlighted the following as key visualization challenges: analysis & comparison of macromolecular ensembles; uncertainty / confidence visualization; mapping of abstract data onto 3D structures, including text, URLs, community curations, as well as data from networks, pathways, populations, geographic distributions, and phylogenies. In accordance with the previous goal of critical assessment, it would help to define several concrete showcases: an example could the 3D models of HIV being developed by Johnson et al. [2], as these combine many of the data types mentioned above.

Education. Science educators regularly employ data visualization techniques to communicate structural biology, but many biological systems have well established visual representations that conflict or entirely disregard best practices identified by data visualization practitioners. Communication of best practice is essential in order to ensure that new approaches for mesoscale structural visualization maximize the potential of these visual analysis tools, and correct terminology employed to allow biologists to discuss data visualization approaches with computer scientists.

Conclusions

Structural biology can now provide a detailed view of the information environment of systems biology. As the available data on structures and omic-scale systems grow and become more complex, the visualization tools developed in both communities need to be integrated with datavis methods into new toolsets for enhancing both exploration and understanding of these biological systems.

4.4 Ontologies in Biological Data Visualization

Sheelagh Carpendale, Min Chen, Daniel Evanko, Nils Gehlenborg, Carsten Görg, Lawrence Hunter, Francis Rowland, Margaret-Anne Storey, Hendrik Strobelt

Introduction

Ontologies are graph-based knowledge representations in which nodes represent concepts and edges represent relationships between concepts. Ontologies have been used extensively as computational models in natural language processing, artificial intelligence, and the web sciences. A number of disciplines in which visualization plays an important role, including biology, are using ontologies to support the analysis of large and complex datasets. We examined how ontologies can be used to support biological data visualization and identified challenges and opportunities from the perspectives of three different stakeholders: ontologists (who create and maintain ontologies), data curators (who use ontologies for annotation purposes), and data analysts (who use ontologies through applications to analyze experimental data). A summary of the challenges and opportunities is presented below; we also submitted a more detailed discussion as a Viewpoints article on Ontologies in Biological Data Visualization to the IEEE Computer Graphics & Applications journal.

Challenges

A first challenge is centered around the dynamic nature of ontologies. Many ontologies constantly change and evolve due to discoveries and newly acquired knowledge in the domain they represent. The creation of multiple versions of ontologies is prone to inconsistencies and also adds downstream complexity for their users (humans as well as computer programs). Keeping track of evolution becomes even more daunting for ontologies that integrate multiple data sources. The evolution of ontologies affects all three types of stakeholders.

A second challenge is scale: many ontologies represent an overwhelming amount of data. These large ontologies are usually developed and maintained by a team of ontologists which requires a framework that supports the collaborative work on ontologies. Data curators often face the problem of finding the most appropriate concepts in these large ontologies when they annotate terms in documents or samples in experimental data. For data analysts, the amount of ontology annotations can be easily overwhelming, especially if documents or data are annotated with multiple ontologies.

A third challenge is related to the relationships and types that are represented in ontologies. So far, most applications do not take advantage of the complex set of relationships in ontologies but rather reduce them to a simple hierarchy. While this approach is certainly useful for data analysts it does not exploit the full potential of ontologies. The underlying problem here is that the representation and visualization of complex relationships is hard. The complex set of relationships within ontologies, combined with the large size of ontologies, makes their manual maintenance a considerable effort for ontologists

Finally, to make ontologies more useful for data analysis, it is crucial to understand what analysts want to investigate and how they can use ontologies for their specific tasks. To this end, user studies are required to understand the workflows and aims of the analysts.

Research Opportunities

Visualization of Ontologies: Despite much research on the topic of ontology visualization, the majority of the tools developed thus far are focused on visualizing or navigating the ontologies themselves, rather than on visualizing content that has been annotated with ontological concepts. We propose that there is a need to develop tools that are both powerful and easy to use for curators of content as well as for users browsing ontologically annotated content (such as journal articles). Such tools can furthermore support the consumers of this content to do richer analyses of associated content.

Automated Generation of Visualization using Ontology: The availability of domain-specific ontologies provides an exciting opportunity for developing automated visualization methods and services. Although interaction remains as an important apparatus for facilitating data exploration, it may incur costly time and learning effort for using a visualization system. In many application scenarios, automatically-generated visualizations may serve users more efficiently and effectively, and can facilitate knowledge sharing among users.

Visualization of Ontological Context in Supporting Search: The application of ontologies and even multiple ontologies to annotate text corpora offers new potential and new challenges. Search is currently being explored within ontologies. This can be expanded to consider search across multiple related ontologies and inverted to include search via multiple ontologies. Ontologically annotated text provides semantically rich meta-data. Since visualizing even simple meta-data has been shown to enhance serendipity in information exploration, visualizing ontologically annotated text is very promising. However, the complexity of text visualization coupled with the complexity of ontology visualization makes this a big challenge.

Conclusion

By capitalizing on ontologies as knowledge representations, we will be able to make a significant step towards the realization of knowledge-assisted visualization [3]. It may take the form of automated visual annotation of texts, documents and corpora, automated construction of visualization for novice users, or automated visualization of ontological context in information retrieval.

4.5 A Framework for Effective Visualization Design

Miriah Meyer, Jan Aerts, Dan Evanko, Jean-Fred Fontaine, Martin Krzywinski, Raghu Machiraju, Kay Nieselt, Jos Roerdink, and Bang Wong

Visualizations are not only an important aspect of how scientists make sense of their data, but also how they communicate their findings. The techniques and guidelines that govern how to design effective visualizations, however, can be quite different whether the goal is to *explore* or to *explain*. For example, if the goal is to support hypothesis generation from a large, genomics data set then techniques like multiple-linked views and data-rich visual representations are good considerations, as opposed to a visualization used in a conference presentation where significant abstraction of the data and simple visuals are necessary.

Unfortunately, scientists are often not aware of the spectrum of considerations when creating visualizations, resulting in ineffective figures, diagrams, and tools when there is a mismatch between the goal of the visualization and the design decisions. These considerations encompass audience, presentation modality, and the amount explanation versus exploration that is needed. To help clarify this problem, we have developed a framework to reason about the spectrum and considerations to help scientists better match their visualization goals with appropriate design considerations. We believe that awareness about this spectrum can improve visualization, particularly those targeting explanatory goals, as well as enable more fruitful discussions between scientists and visualization designers.

The framework (Figure 7) has a major axis that describes how exploratory or explanatory a a visualization task is. On one end, pure exploratory visualizations are mostly likely to be interactive, and are often meant to support hypothesis generation, data and model validation, and scientific insight. On the other end, pure explanatory visualization are meant to communicate an idea, story, or scientific finding, usually in a highly abstract, simple way. It is important to map goals to a position along this axis because different design considerations exist from left to right. Things to consider are: who am I communication to? Someone in my lab? In my department? In my scientific community? In the general public?

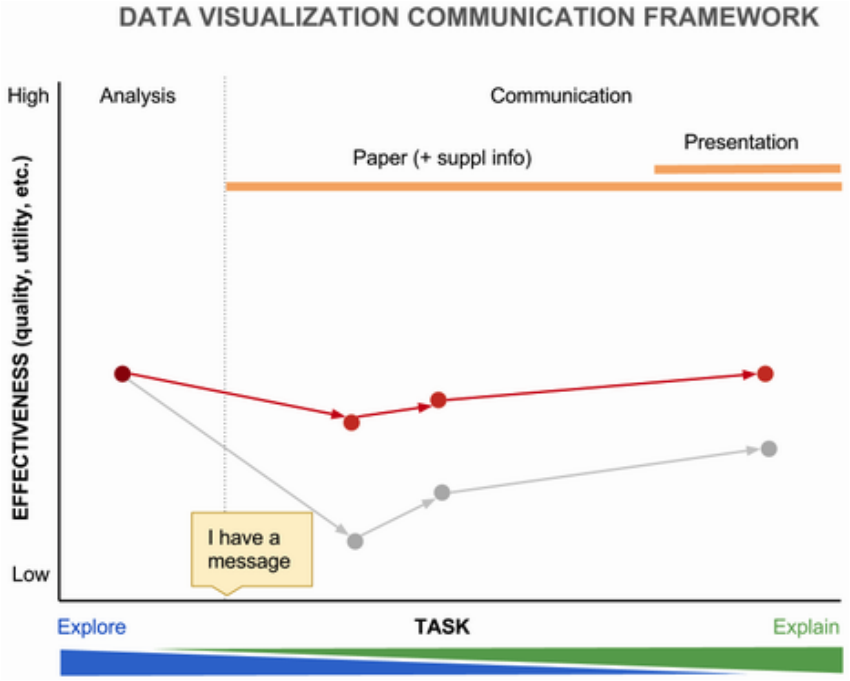


Figure 7 Data Visualization Communication Framework.

These different locations on the task axis have different design consideration. For exploratory visualizations, these considerations are largely drawn from computer science, and are things such as interactivity, how to display or summarize large data sets, and how to support complex relationships. For explanatory visualizations, these considerations come largely from the design community, such as how to abstractly represent data and relationships, how to filter out unnecessary data and details, and how to tell a story visually.

The task axis also coincides with numerous other secondary consideration axes. Some of these are: considering the richness of the data, how much complexity can be shown; considering the amount of data, how much of the data must be filtered out; considering

hypothesis generation, can the viewer form new and different hypotheses; considering the time commitment of the viewer, can they get the message in 5 seconds, 5 minutes, or 5 hours; and considering the domain expertise, how much specific knowledge does it require.The second axis describes the effectiveness of the visualization. Considering this axis there is an important implication: moving back and forth has gravity. What this means is that an effective visualization at one point within the space will almost always be less effective when used directly for an application further along the major axis. For example, using an interactive genome browser with a full data set is effective for exploration, but will perform terribly in a conference presentation on a scientific finding. And conversely, a diagram for the general public explaining a scientific concept will almost certainly fail to produce new hypotheses for a research scientist. In summary, knowing where you are within the framework can help in picking appropriate design guidelines and visualizations for a specific communication goal.

4.6 Uncertainty Visualization

Min Chen, Julian Heinrich, Jessie Kennedy, Andreas Kerren, Falk Schreiber, Svenia Simon, Christian Stolte, Corinna Vehlow, Michel Westenberg, and Bang Wong

Uncertainty is common in all areas of science, and it poses a hard problem to deal with in terms of visualization. There is no well-established definition of uncertainty, but several types of uncertainty are generally distinguished: measurement precision, completeness, inference, credibility, and disagreement [18]. Visualization of uncertainty has received much attention in the areas of scientific visualization and geographic visualization. Several techniques have been proposed employing special visual encodings (transparency, blur, error bars), addition of glyphs, modification of geometry, and animation, to name a few. Application of uncertainty visualization appears much less common in information visualization and in biological data visualization.

The working group looked at sources of uncertainty and types of uncertainty specific to biology. We studied a (RNA) sequencing (RNAseq [21]) pipeline, and identified the types of ambiguity that can be introduced in each step, see Fig. 8. Many steps are prone to introduce errors, some of which create a certain bias in what RNA fragments are ultimately amplified and sequenced. The output that comes from the pipeline is a set of mapped reads with an associated quality value that could be used (but is rarely in practice) in visualizing the sequences.

We also looked at computational models derived from the literature. Here, uncertainty is apparent in the model itself (granularity and structure), in the simulation (initial parameters, numerical inaccuracy), and verification of the simulated model by lab experiments (measurement errors). An open issue that we identified here concerns uncertainty visualization in networks (which represent the model): uncertainty in network topology, and the problem of dealing with dynamic (uncertain) attributes on nodes, edges, and their interdependencies.

The working group performed a quick scan of recent papers that employ some form of uncertainty visualization for biological data. We found several examples, including applications in population variability [4], expression data analysis [9, 22], data cleansing [15], and network modeling [16, 20].

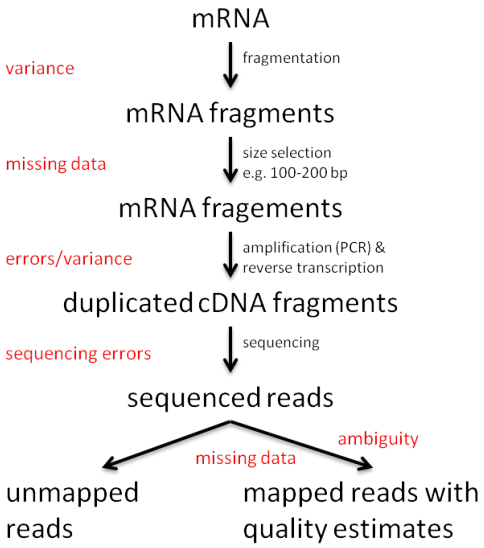


Figure 8 Uncertainty (red text) in the sequencing pipeline.

Our plan is to further extend the result of this quick scan into a literature survey paper that specifically addresses uncertainty visualization in biology. We propose to construct a taxonomy of uncertainty visualization approaches, and investigate how they could be employed in the context of a collection of biological problems.

4.7 Evaluation

Jarke J. van Wijk, Kasper Dinkla, Martin Graham, Graham Johnson, Francis Rowland, and Andrea Schafferhaus

In our breakout group we discussed issues concerning evaluation in biovis. We agreed that two problems were central. First, how to get biologists motivated to participate in evaluations? During the talk we had a questionnaire send out and asked for feedback. The main result was (simply) that biologists have to get something out of it to motivate them: a tool they can use to solve their problems, and also chocolate and beer were suggested.Second, we focussed on how to evaluate. We discerned a number of dimensions, centered around what, why, when, where, and how. We found a strong difference between designers working at a bio-institute and infovis researchers. The former is characterized by close cooperation during development, a qualitative approach, and a focus on creating a tool; whereas the latter group performs evaluation typically afterwards, aims at quantitative results and creation of new techniques. Both have merit, the challenge is to close the gap.We decided that it would be great to develop a tool that shows various options for evaluation methods, given a specification of the problem, and designed a first mock-up, consisting of a set of filters/choices and a scatterplot showing approaches (see Figure 9). We aim to develop this idea further after the seminar.

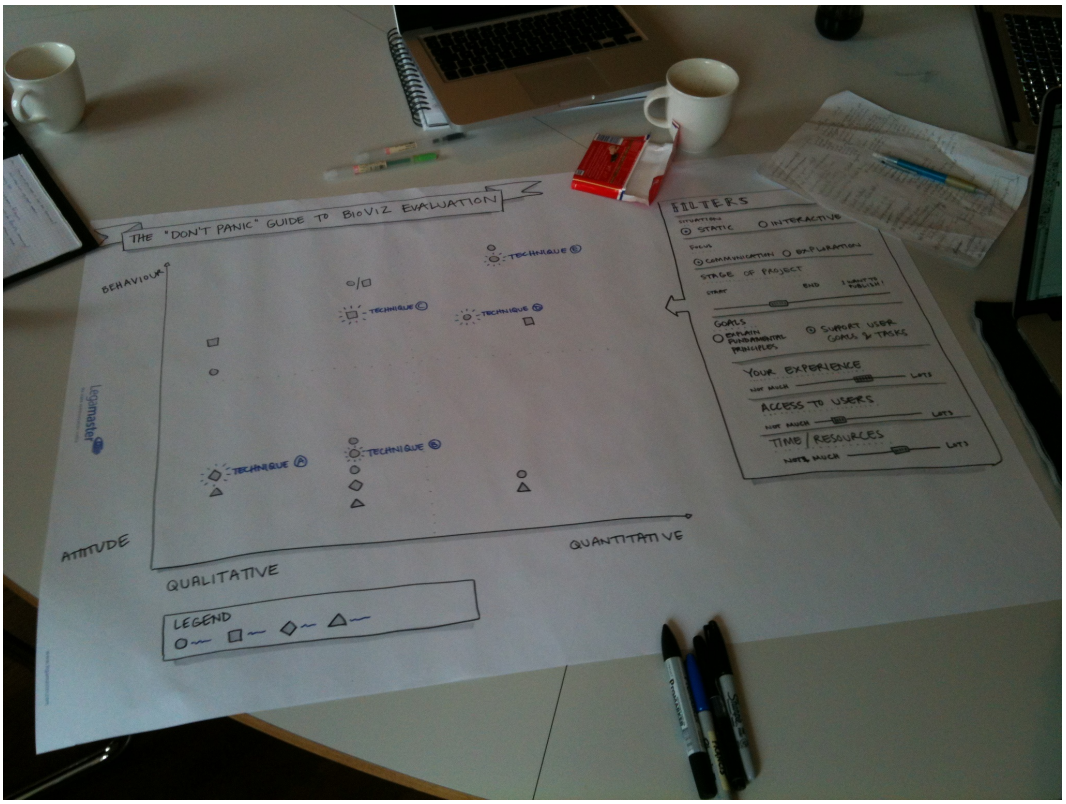


Figure 9 First sketch of a tool that shows various options for evaluation methods.

4.8 Multiscale Visualization

Carsten Görg, Graham Johnson, Karsten Klein, Oliver Kohlbacher, Thorsten Möller, Arthur Olson, Francis Rowland, and Matt Ward

Overview

In recent years the progress in the understanding of biological processes, in combination with the corresponding collection of large amounts of heterogeneous experimental data, led to raised requirements for corresponding visualizations. These visualizations should allow biologists to analyze data with respect to complex and even whole-organism models [10], that combine data of differing type, multiple dimensions such as space and time, as well as multiple scales in those dimensions.

While there are still visualization problems to solve for single dimensional data (e.g. effective comparison or representation of dynamics for protein interactions changing over time), the additional challenges here include the linking of different scales and types of data for presentation or interactive exploration. Stephen Prevenas (thelazygeeks.com) uses the following analogy to describe the shortcomings of just combining current single scale solutions:

Think of trying to watch The Empire Strikes Back, but the actors are on your TV,

> the scenery is on your iPad, and the soundtrack is on an 8-track. Ok, maybe not an 8-track, that makes this analogy absurd. Say the soundtrack is on a CD, which I admit is only marginally more believable at this point, but I also shouldn't assume you have an iPad and an iPod.

We started to study these challenges, identify the underlying problems, and summarize them to lay the foundation for further research.

Goals and Discussion

We first had an exchange on the personal background and expectation of the group members regarding the working group. Even though the backgrounds were quite heterogeneous, including people from visualization, design, bioinformatics, and biology, the expectations were similar. We agreed that multiscale visualizations of biological data have many aspects and cover such diverse application fields that we first had to agree on a basic characterization of what multiscale means in the context of our discussion before we could even start working on the corresponding challenges. In addition, we wanted to collect examples of practical multiscale problems that combine multiple dimensions of scale, which could help us to derive a characterization.

We thus worked towards the following two aims:

1. Discussing and defining a formal characterization of multiscale visualizations of biological data.
2. Collecting multi-scale visualization examples to foster our search for a practically useful characterization and to present to the seminar participants.

It turned out that both problems were not easy to solve. We did not find examples that cover more than one or two dimensions (usually spatio-temporal dimensions), and several levels of scale; in fact, we were not aware of well-defined levels and dimensions besides the standard spatial and temporal dimensions. We also agreed that there is no sufficient, and at the same time unambiguous and generally accepted definition of multiscale in the domain of biology, and that this lack of definition would hinder further discussions in our group. We therefore decided that the first task of our group would be to develop our own definition, or at least try to cover the most important aspects in a characterization, and that we then should discuss the challenges and open problems with respect to the most interesting tasks and data.

The main questions we discussed in the following tried to link these two tasks: How can multiscale examples be systematically classified or categorized and what are reasonable dimensions in which scaling takes place. We spent half of the time of our meeting to discuss the nature of model and data modalities, different corresponding potential dimensions, and how they could be clearly separated. It turned out that it was a difficult task to agree on a simple but precise characterization that is useful as a base for further investigation. As our first approach to come to some understanding what defines a multiscale visualization challenge, we decided to describe corresponding processes, data, and tasks in terms of a coordinate system that is made up of the dimensions that fully characterize their multiscale nature. However, we discovered that the proposed dimensions often overlapped, were impossible to grasp and define formally, or seemed not to be useful enough for further investigations. In particular, there was a long discussion about the term "data complexity". The questions we asked were, among others:

- How can "complexity" of data be captured, is it represented e.g. by the entropy, and does it include the size of the data?

- Is instead data size one of the scales, and if yes, does this cover problems where different data sizes are part of the visualization output, e.g. for comparison, or also cover problems where different amounts of input data are processed?

Obviously data volume does not directly translate into content information, but content information alone does not fully cover the corresponding problems. In the end, we decided that due to problems like limited human perception and computational complexity the data volume, i.e. the data size, has to be taken into account as one of the data dimensions. Since we also wanted to capture at least some aspect of complexity besides data volume, we discussed to also add the number of sources of the data, i.e. the method to generate it, and the number of representations as additional dimensions. In the end we reduced our characterization to an easy to understand four dimensional coordinate system that covers the complexity at least to a significant extent. These four dimensions of scaling include the quite natural and well-accepted time and space dimensions. In addition, we chose the number of modalities, which is a way to cover complexity without having to define a complexity metric for all kinds of data, and data volume as the two other dimensions.

After we agreed on this basic characterization we started collecting examples of multiscale visualizations to (1) provide an understandable view of the nature of the different scales and how they apply to biological data and exploration, and (2) to investigate the shortcomings of existing approaches. As a goal, we would like to bring together good and bad examples, where a classification according to the extent with which they cover the dimensions of our multi-scale characterization should allow better access to our collection.

We decided to produce a short informational video that presents multiscale visualizations for the different dimensions in our coordinate system, both for presentation and getting feedback from the seminar participants, as well as a prototype for a result to be published later. To have a clear and easy to understand demonstration of the different dimensions, we selected a common topic that allows to cover each of them. We chose the human body for that purpose and picked examples for each dimension that represent different aspects.

Another (short) topic of discussion was how to actually represent different scales like cellular and molecular level. Several approaches exist: (1) fly through the scales consistently from one to another, (2) combined visualizations (focus), and (3) multiple linked views (e.g. a magnifying glass). Linked to that is the problem of transition between different modalities (like a graph and the physical representation), and if they can be combined or need to be visualized in parallel. In order to make use of these representations, suitable interaction techniques need to be developed, but this discussion was outside the scope of this working group.

Outcome

"Understanding Multi-Scale Visualization" is a design and prototype for a short video exposition on the nature of multi-scale data, models and visualization in biology. Its goal is to provide a quick and understandable view of the nature of the different scales and how they apply to biological data and exploration. Our characterization could be used as a foundation for a taxonomy of multiscale techniques, and our example collection could show which parts of biological data space have been explored already.

4.9 Infrastructure

Seán O'Donoghue, Tom Freeman, Mark Ragan, Margaret Storey, Larry Hunter, Cydney Nielsen, Nils Gehlenborg, Jim Procter, and Hendrik Strobelt

Overview

The infrastructure working group discussed needs from both a technical and community standpoint regarding the challenges involved in the analysis of biomedical data derived from the Cancer Genome Atlas project and mechanisms to facilitate interactions between visualization communities in computer science and biology. Eight key criteria were identified: *Interoperability*, *reusability*, *compatibility*, *references & benchmarks*, *middleware*, *vertical integration*, *scalability*, and *sustainability*, and two outcomes. The first outcome is a model for a community-maintained, biological visualization resource that would enable biological questions, task descriptions, sample datasets and existing tools for the problems to be disseminated to the computational visualization and biological research communities. The second is a detailed use-case based on the data and analysis pipelines of the cancer genome atlas that will allow technical aspects of the eight key criteria to be explored and practical solutions proposed.

Key Criteria for BioVis Infrastructure

Interoperability. The success of systems such as Galaxy [5] and Vistrails (http://www.vistrails.org) demonstrates that BioVis tools developed by different groups in the community must interoperate, at the very least through consistent data exchange standards, but also in the provision of well designed and documented control interfaces to allow pipelining and orchestration.

Reusability. Best practices are needed to encourage groups to develop tools with standard interfaces that allow them to be embedded or combined with other tools (e.g. as widgets) in a variety of situations.

Comparability. Two aspects were identified: It should be straightforward to compare different tools that perform a similar function in order to assess which one is most appropriate for a use case. Effective BioVis tools should also provide visualizations that support comparative analysis of biological data.

References & Benchmarks. A standard model and repository is needed to allow reference biological problems to be described, along with representative datasets and analysis outcomes. Benchmark problems should be significant and representative of key biological questions, and task descriptions should be presented in a predigested manner such that a non-specialist in the field can understand the analysis processes required without deep knowledge. These reference descriptions can be used a benchmarks to evaluate existing BioVis tools and inform the design of new tools in the future.

Middleware. Middleware can facilitate *interoperability*, *reusability*, *comparability*, and *vertical integration*, providing it is easy to use, it also lowers the cost of entry to the field. In fact, several middleware libraries exist for biological data, but they are typically associated with a particular modality or biological domain. A new breed of middleware is needed that can draw together standard technical solutions from all relevant biological information

domains, such as image processing, text mining, and biomolecular sequence and structure analysis.

Vertical integration. Computational analysis pipelines are commonly used in biology, but it is essential that these can be published in a way that allows an analysis to be reproduced by other researchers. A number of systems that support provenance and pipeline dissemination have been developed in computer science (e.g. myExperiment, VisTrails, etc.) but the Bio & Vis communities must work to make their use routine in data driven biology.

Scalability. It is safe to assume that any notion of 'big data' currently described will be relatively small compared to the volume of data that must be handled by our tools in the future. Technical solutions (e.g. data/processing clouds) already exist, but the users of the system are often the most serious bottleneck and data and derived analysis results need to be delivered in a usable manner.

Sustainability. Open Sustainability models are required for software as well as data created by grant-based biological research, to ensure the community at large can access the outcomes of research. Such practices are not commonplace in computer science laboratories, where prototypes serve only as proof-of-concepts to be abandoned rather than refined to make them usable by biologists. New standards for software tools must be declared in both communities, and public repositories (such as the one described below) should be created to enable rapid interchange of new tools and datasets, and maintenance of previously developed tools.

Outcomes

We decided to develop two outcomes following our initial discussions. A community resource for biological task, data and tool dissemination and a case study to explore the software infrastructure necessary to support a current biomedical research problem.

Community resource for benchmark problems, datasets and available tools
A web platform to support community maintained descriptions, datasets and instances of tools relevant to BioVis could be provided that would act as a bridge between the CS and biological BioVis communities by integrating with both the biovis.net and vizbi.org sites. It will provide:

- Biological problem/analysis task descriptions described in a way that is accessible to lay-biologists and computer scientists.
- Data sets relevant to problems, using standard file formats or links to archive quality web based databases.
- Descriptions of available tools. Each tool should be linked to, or archived on the site so that it can be launched, along with instructions describing how to perform the task with this tool.

It is essential that tasks, datasets, and tools provided by the resource are *significant*, *representative*, *selective*, and *predigested* in order to ensure that the resource is relevant to both the computer science and biological community. A number of starting points were proposed, including reaping problem descriptions and tools that solve problems from the burgeoning number of Stack Overflow style sites [19] and deriving descriptions from the reviews created by the VIZBI community. In order to be sustainable the resource will require a community of editors to be recruited and the engagement of tool authors to maintain the public descriptions of their work.

Use case based on the Cancer Genome Atlas (CGA)
"The CGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing." (from the CGA's own materials). Although at a preliminary phase, the numbers involved in this experimental study are already staggering: 20 tumor types are being investigated by analyzing affected and unaffected tissue from 500 patients to identify variation in single nucleotide polymorphisms (SNPs), copy number (CNVs), DNA methylation, mRNA and microRNA transcripts, and gene mutations.

Data from each experiment requires one or more computational analysis pipelines, and the results must be validated, integrated and understood in order to elucidate the driving mechanisms in each tumor type, and evaluate the efficacy of available therapies for each individual. The CGA have developed a data and result staging system, *Firehose*, and a visualization tool – *StratomeX*, based on *Caleydo* (an Eclipse Rich Client Platform (RCP)), that provides genome-scale integrated genome/transcriptome views across these data. However, a number of data curation and deep biological analysis tasks are very difficult.

The following criteria were identified for a next generation CGA visualization system:

1. Data Provenance. System needs to display the origin and processing pathway for the data currently visualized, and ideally allow comparison of results of alternate processing pathways.
2. Standardized representation of data and analysis processes. Provenance models require well defined representations such as the Predictive Model Markup Language [Wikipedia] to describe the transformations that data undergoes prior to visualization. Formal representations also enable interoperability and reproducibility.
3. Remote access to data. Most sets of data are too large to fit into memory – the system needs aggregation and subsetting mechanisms to allow browsing of the complete dataset on any reasonable user platform.
4. Pluggable architecture. Alternate visualizations for the same data, or additional visualizations for new or derived data facilitate deep analysis, and encourages contribution from third-parties. Architecture will also allow new data format and analysis process support.
5. Communication between plugins/modules. Synergies are important: communication between distinct visualization modules with shared selections, colorings, etc. allow greater insight. The modules in the system should also be able to select appropriate modules for performing particular purposes – such as computing a distance matrix for particular biological entities and then clustering to yield an appropriate visualization.
6. Global communication/user experience. Several different types of users will enter and use the system in different ways – a core system event model will need to support arbitrary routes through the data and analysis process.

These criteria were explored further, to identify base layer software components and the key visualizations that would be needed and the kind of provenance associated with each one. A user story describing how a typical cancer biologist will employ the system was proposed to explore how the different analysis and visualization components will be required to interact.

Conclusion

The infrastructure working group identified a number of technical and social requirements that should be addressed by the community. We proposed the development of a community resource to collate problems and solutions for biological visualization tasks, and this will be explored further. We also developed a detailed use case based on a current biomedical research

problem to help identify technical and conceptual challenges in biological visualization that the community should prioritize in the future.

5 Acknowledgements

We would like to thank all participants of the seminar for their contributions and lively discussions; we also would like to thank the reviewers of our initial proposal for their constructive feedback and the scientific directorate of Dagstuhl Castle for providing us with the opportunity to organize this seminar. Finally, the seminar would not have been possible without the untiring help of the (scientific) staff of Dagstuhl Castle, including Ms. Susanne Bach-Bernhard, Ms. Jutka Gasirorowski, and Dr. Marc Herbstritt.

References

1 Jan A Aerts and Chris Tyler-Smith. *Structural Variation in Great Ape Genomes*. John Wiley & Sons, Ltd, 2001.
2 autoPACK model of HIV 1.4 running in PMV with Screen Space Ambient Occlusion Narrated. http://www.youtube.com/watch?v=W84yW9HIzCI.
3 Min Chen, D. Ebert, H. Hagen, R.S. Laramee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *Computer Graphics and Applications, IEEE*, 29(1):12 –19, jan.-feb. 2009.
4 M. Corell, S. Ghosh, D. O'Connor, and M. Gleicher. Visualizing virus population variability from next generation sequencing data. In *Proc. IEEE Symp. Biological Data Visualization (BioVis)*, pages 135–142, 2011.
5 J. Goecks, A. Nekrutenko, J. Taylor, E. Afgan, G. Ananda, D. Baker, D. Blankenberg, R. Chakrabarty, N. Coraor, J. Goecks, G. Von Kuster, R. Lazarus, K. Li, A. Nekrutenko, J. Taylor, and K. Vincent. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11(8):R86, 2010.
6 D. Gray. *Gamestorming: A Playbook for Innovators, Rulebreakers, and Changemakers*. O'Reilly, 2010.
7 A. Herbig, G. Jager, F. Battke, and K. Nieselt. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, 28(12):7–15, Jun 2012.
8 L. Hohmann. *Innovation Games: Breakthrough Products Through Collaborative Play: Creating Breakthrough Products and Services*. Addison Wesley, 2006.
9 C. Holzhüter, H. Schumann, A. Lex, D. Schmalstieg, H.-J. Schulz, and M. Streit. Visualizing uncertainty in biological expression data. In *Proc. SPIE 8294, Visualization and Data Analysis (VDA 2012)*, 2012.
10 J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival, N. Assad-Garcia, J. I. Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, Jul 2012.
11 M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res.*, 19(9):1639–1645, Sep 2009.
12 C. Nielsen and B. Wong. Points of view: Representing genomic structural variation. *Nat. Methods*, 9(7):631, Jul 2012.
13 C. B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, and T. Wang. Visualizing genomes: techniques and challenges. *Nat. Methods*, 7(3 Suppl):S5–S15, Mar 2010.
14 B. Paten, M. Diekhans, D. Earl, J. S. John, J. Ma, B. Suh, and D. Haussler. Cactus graphs for genome comparisons. *J. Comput. Biol.*, 18(3):469–481, Mar 2011.

15 T. Paterson, M. Graham, J. Kennedy, and A. Law. VIPER: a visualisation tool for exploring inheritance inconsistencies in genotyped pedigrees. *BMC Bioinformatics*, 13(Suppl 8):S5, 2012.

16 H. Rohn, A. Hartmann, A. Junker, B. H. Junker, and F. Schreiber. FluxMap: a Vanted add-on for the visual exploration of flux distributions in biological networks. *BMC Systems Biology*, 6:33, 2012.

17 A. Roulin, P. L. Auer, M. Libault, J. Schlueter, A. Farmer, G. May, G. Stacey, R. W. Doerge, and S. A. Jackson. The fate of duplicated genes in a polyploid plant genome. *Plant J.*, Sep 2012.

18 M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualization. In *Proc. AVI'08*, pages 376–379, Napoli, Italy, 28–30 May 2008.

19 Margaret-Anne Storey, Christoph Treude, Arie van Deursen, and Li-Te Cheng. The Impact of Social Media on Software Engineering Practices and Tools (DCS -338-IR). In *FSE/SDP Workshop on the Future of Software Engineering Research (FOSER 2010)*, pages 359–364, 2010.

20 C. Vehlow, J. Hasenauer, A. Kramer, J. Heinrich, N. Radde, F. Allgöwer, and D. Weiskopf. Uncertainty-aware visual analysis of biochemical reaction networks. In *Proc. IEEE Symp. Biological Data Visualization (BioVis 2012)*, volume 2012, pages 91–98, 2012.

21 Z Wang, M Gerstein, and M Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

22 M. A. Westenberg, J. B. T. M. Roerdink, O. P. Kuipers, and S. A. F. T. van Hijum. SpotXplore: a Cytoscape plugin for visual exploration of hotspot expression in gene regulatory networks. *Bioinformatics*, 16(22):2922–2923, 2010.

Participants

- Jan Aerts
K.U. Leuven, BE
- Sheelagh Carpendale
University of Calgary, CA
- Min Chen
University of Oxford, GB
- Kasper Dinkla
TU Eindhoven, NL
- Daniel Evanko
Nature Publishing Group, US
- Jean-Fred Fontaine
Max-Delbrück-Centrum, DE
- Tom Freeman
University of Edinburgh, GB
- Nils Gehlenborg
Harvard University, US
- Carsten Görg
University of Colorado, US
- Martin Graham
Edinburgh Napier University, GB
- Julian Heinrich
Universität Stuttgart, DE
- Lawrence Hunter
University of Colorado, US
- Graham Johnson
UC – San Francisco, US
- Jessie Kennedy
Edinburgh Napier University, GB
- Andreas Kerren
Linnaeus University – Växjö, SE
- Karsten Klein
The University of Sydney, AU
- Oliver Kohlbacher
Universität Tübingen, DE
- Martin Krzywinski
BC Cancer Research Centre, CA
- Michael Lappe
CLC bio, DK
- Raghu Machiraju
Ohio State University, US
- Miriah Meyer
University of Utah – Salt Lake City, US
- Torsten Möller
Simon Fraser University – Burnaby, CA
- Cydney Nielsen
BC Cancer Agency's Genome Sciences Center, CA
- Kay Nieselt
Universität Tübingen, DE
- Sean O'Donoghue
CSIRO – North Ryde, AU
- Arthur J. Olson
The Scripps Research Institute – La Jolla, US
- James Procter
University of Dundee, GB
- Mark Ragan
The Univ. of Queensland, AU
- Jos B.T.M. Roerdink
University of Groningen, NL
- Francis Rowland
EBI – Cambridge, GB
- Andrea Schafferhans
TU München, DE
- Falk Schreiber
IPK Gatersleben, DE
- Svenja Simon
Universität Konstanz, DE
- Christian Stolte
CSIRO – North Ryde, AU
- Margaret-Anne Storey
University of Victoria, CA
- Hendrik Strobelt
Universität Konstanz, DE
- Jarke J. Van Wijk
TU Eindhoven, NL
- Corinna Vehlow
Universität Stuttgart, DE
- Matthew O. Ward
Worcester Polytechnic Inst., US
- Michel A. Westenberg
TU Eindhoven, NL
- Bang Wong
Broad Institute of MIT & Harvard – Cambridge, US

Report from Dagstuhl Seminar 12381

Privacy-Oriented Cryptography

Edited by

Jan Camenisch[1], Mark Manulis[2], Gene Tsudik[3], and Rebecca N. Wright[4]

1 **IBM Research – Zürich, CH,** jca@zurich.ibm.com
2 **University of Surrey, GB,** mark@manulis.eu
3 **University of California – Irvine, US,** gts@ics.uci.edu
4 **Rutgers University, US,** rebecca.wright@rutgers.edu

Abstract

This report documents the program of the Dagstuhl Seminar 12381 "Privacy-Oriented Cryptography", which took place at Schloss Dagstuhl in September 16-21, 2012. Being the first Dagstuhl seminar that explicitly aimed to combine cryptography and privacy research communities, it attracted a high number of participants, many of whom were new to Dagstuhl. In total, the seminar was attended by 39 international researchers, working in different areas of cryptography and privacy, from academia, industry, and governmental organizations. The seminar included many interactive talks on novel, so-far unpublished results, aiming at the design, analysis, and practical deployment of cryptographic mechanisms for protecting privacy of users and data. The seminar featured two panel discussions to address various approaches towards provable privacy and different challenges but also success stories for practical deployment of existing cryptographic privacy-oriented techniques.

Seminar 16.–21. September, 2012 – www.dagstuhl.de/12381
1998 ACM Subject Classification D.4.6 Security and Protection
Keywords and phrases Privacy, Cryptography, Anonymity, Confidentiality
Digital Object Identifier 10.4230/DagRep.2.9.165

1 Executive Summary

Jan Camenisch
Mark Manulis
Gene Tsudik
Rebecca N. Wright

The constantly increasing volume of electronic interactions and sensitive information disseminated online raises privacy concerns and motivates the need for efficient privacy-oriented techniques. The aim of our "Privacy-Oriented Cryptography" seminar was to bring together (mainly, but not only) researchers working in different domains of cryptography and privacy. Although non-cryptographic measures can, at times, aid privacy (e.g., statistical or ad hoc obfuscation techniques) — cryptography, via its mathematical mechanisms and formal concepts, helps obtain novel and efficient privacy-enhancing solutions, achieving concrete and measurable privacy guarantees.

Since privacy is a very broad area, being explored not only by security and cryptography experts, this seminar focused on two domains: *user privacy* and *data privacy*, for which the

benefit from using cryptographic techniques is especially significant. Seminar participants presented and discussed many novel privacy-oriented cryptographic algorithms and protocols that admit various fields of deployment for protecting privacy in a broad range of applications, involving possibly huge amounts of data (e.g., cloud computing) and many different users (e.g. online communities). The seminar further addressed the emerging research direction of *provable privacy*, by discussing various mechanisms and techniques for defining concrete privacy goals and enabling their formal analysis.

The seminar brought together 39 of the leading scientists in the areas of (applied) cryptography and privacy. The participants came from all over the world, including the US (13 participants), Germany (8), Switzerland (6), Great Britain (5), Australia (1), Belgium (1), Canada (1), France (1), Italy (1), and Sweden (1).

The program contained 26 interactive presentations, each about 35–40 minutes and two panel discussions, with a free afternoon on Wednesday to offer time for social activities or for conducting collaborative research in smaller groups. The seminar ended on Friday after lunch to enable time for traveling. We asked participants prior to the seminar to suggest talks based on their most recent results. Most presentations followed this suggestion and introduced new, sometimes even not yet submitted or still work-in-progress results. The first panel — "Privacy Models: UC or Not UC?" — discussed the advantages and disadvantages of existing cryptographic methods for formal specification and analysis of security and privacy guarantees. The second panel — "Privacy-Oriented Cryptography: Why is it not adopted more in practice?" — discussed challenges that arise in the practical deployment of existing privacy-oriented cryptographic solutions but also considered some success stories like Tor, a popular anonymous communications service, which is widely used in different parts of the world.

The nature of the seminar allowed experts and practitioners to air ideas and discuss preliminary concepts and work-in-progress results. This might have led to the exposure and subsequent exploration of new research directions that may offer both practical significance and intellectual challenge.

The organizers would like to thank all participants for accepting our invitations and attending the seminar, and for sharing their ideas and contributing to the interesting seminar program. We hope that discussions were fruitful and the opportunity to work face-to-face during the seminar helped to create impulses for exciting new research projects, paving the way for further progress and new discoveries in Privacy-Oriented Cryptography.

Finally, the organizers, also on behalf of the participants, would like to thank the staff and the management of Schloss Dagstuhl for their support throughout the 1,5 years of preparations of this very pleasant and successful event.

2 Table of Contents

3 Overview of Talks

3.1 Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers

Giuseppe Ateniese (Johns Hopkins University – Baltimore, US)

Joint work of Ateniese, Giuseppe; Felici, Giovanni; Mancini, Luigi V.; Spognardi, Angelo; Villani, Antonio; Vitali, Domenico

Machine Learning (ML) algorithms are used to train computers to perform a variety of complex tasks and improve with experience. Computers learn how to recognize patterns, make unintended decisions, or react to a dynamic environment. Certain trained machines may be more effective than others because they are based on more suitable ML algorithms or because they were trained through superior training sets. The distinctive ability of a particular ML model gradually develops during the training phase, when it is fed with samples of data to form its knowledge. Although ML algorithms are known and publicly released, training sets may not be reasonably ascertainable and, indeed, may be guarded as trade secrets.

While much research has been performed about the privacy of the elements of training sets, in this paper we focus our attention on ML classifiers and on the statistical information that can be unconsciously or maliciously revealed from them. We show that it is possible to infer unexpected but useful information from ML classifiers.

In particular, we build a novel meta-classifier and train it to hack other classifiers, obtaining meaningful information about their training sets. This kind of information leakage can be exploited, for example, by a vendor to build more effective classifiers or to simply acquire trade secrets from a competitor's apparatus, potentially violating its intellectual property rights.

3.2 Attribute-Based Encryption from Lattices

Xavier Boyen (Prime Cryptography – Palo Alto, US)

We initiate the study of a new promising framework for the design of expressive cryptosystems from lattice assumptions. Specifically, we construct the first ever "complex" ABE scheme for a post-quantum world, defeating critical obstacles previously standing in the way of this actively researched result.

3.3 Enabling Complex Queries and RBAC Policies for Multiuser Encrypted Storage

Bruno Crispo (University of Trento – Povo, IT)

Cloud computing has the advantage that it offers companies (virtually) unlimited data storage at attractive costs. However, it also introduces new challenges for protecting the confidentiality of the data. Most current security schemes support an all-or-nothing access model to the data that is too coarse-grained. Moreover, existing schemes do not allow complex encrypted queries over encrypted data in a multi-user setting. Instead, they are limited to keyword searches or conjunctions of keywords.

We extend the work done on multi-user encrypted search schemes by (i) supporting SQL-like encrypted queries on encrypted databases, and (ii) introducing a fine-grained access control over the data stored in the outsourced database based on the RBAC model.

Finally, we implemented our scheme and compare its performance with recent similar solutions.

3.4 Efficient and Secure Testing of Fully-Sequenced Human Genomes

Emiliano De Cristofaro (PARC – Palo Alto, US)

Joint work of Baldi, Pierre; Baronio, Roberta; De Cristofaro, Emiliano; Gasti, Paolo; Tsudik, Gene
Main reference P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, G Tsudik, "Countering GATTACA: efficient and secure testing of fully-sequenced human genomes," in Proc. of the 18th ACM Conf. on Computer and Communications Security (CCS'11), pp. 691–702, ACM, 2011.
URL http://dx.doi.org/10.1145/2046707.2046785

Recent advances in DNA sequencing technologies have put ubiquitous availability of fully sequenced human genomes within reach. It is no longer hard to imagine the day when everyone will have the means to obtain and store one's own DNA sequence. Widespread and affordable availability of fully sequenced genomes immediately opens up important opportunities in a number of health-related fields. In particular, common genomic applications and tests performed in vitro today will soon be conducted computationally, using digitized genomes. New applications will be developed as genome-enabled medicine becomes increasingly preventive and personalized. However, this progress also prompts significant privacy challenges associated with potential loss, theft, or misuse of genomic data.

In this talk, we begin to address genomic privacy by focusing on some important applications: Paternity Tests, Ancestry Testing, Personalized Medicine, and Genetic Compatibility Tests. After carefully analyzing these applications and their privacy requirements, we propose a set of efficient techniques based on private set operations. This allows us to implement in silico some operations that are currently performed via in vitro methods, in a secure fashion. Experimental results demonstrate that proposed techniques are both feasible and practical today.

Finally, we explore a few alternatives to securely store human genomes and allow authorized parties to run tests such that only the required minimum amount of information is disclosed.

3.5 On Anonymity Attacks in the Real-World

Roger Dingledine (The TOR Project, US)

Tor's approach to threat models is to try to understand the capabilities of realistic attackers we expect to encounter, rather than picking adversaries our protocols can withstand. This strategy has led us to deploy systems that are not amenable to security proofs. Or to say it even more strongly, we deploy provably insecure systems relative to real-world adversaries, because they're still the safest ones we can deploy.

In this talk I'll explain some realistic attacks against Tor's anonymity and blocking-resistance properties, and discuss some reasons why it's hard to produce accurate and useful models for these attacks (and thus hard to prove things about them).

3.6 Overcoming Weak Expectations

Yevgeniy Dodis (New York University, US)

Recently, there has been renewed interest in basing cryptographic primitives on weak secrets, where the only information about the secret is some non-trivial amount of (min-)entropy. From a formal point of view, such results require to upper bound the expectation of some function $f(X)$, where X is a weak source in question. We show an elementary inequality which essentially upper bounds such 'weak expectation' by two terms, the first of which is independent of f, while the second only depends on the 'variance' of f under uniform distribution. Quite remarkably, as relatively simple corollaries of this elementary inequality, we obtain some 'unexpected' results, in several cases noticeably simplifying/improving prior techniques for the same problem. Examples include non-malleable extractors, leakage-resilient symmetric encryption, seed-dependent condensers and improved entropy loss for the leftover hash lemma.

3.7 Cryptographic Protocols for Privacy Preserving Access Control in Databases

Maria Dubovitskaya (IBM Research – Zürich, CH)

Joint work of Camenisch, Jan; Dubovitskaya, Maria; Neven, Gregory

We present cryptographic protocols that allow querying a database in a privacy preserving way under access control restrictions. We first design a protocol for the oblivious transfer of data with access control mechanisms. This work can be extended for a practical application of buying records from a database privately with unlinkable priced oblivious transfer protocol. Other extensions use attribute-based encryption and anonymous credentials for obliviously querying a database with hidden access control policies of different complexity.

3.8 Privacy Concepts in the New German Electronic Identity Cards

Marc Fischlin (TU Darmstadt, DE)

We review the privacy aspects of the protocols in the new German electronic identity cards, which have been issued since November 2011. Some of the protocols have also been adopted by the International Civil Aviation Organization (ICAO) and can be expected to be deployed for international machine readable travel documents.

3.9 ZQL: A Cryptographic Compiler for Processing Private Data

Cedric Fournet (Microsoft Research UK – Cambridge, GB)

Joint work of Danezis, George; Fournet, Cedric; Kohlweiss, Markulf; Luo, Zhengqin

Our goal is to enable programming on sensitive data without disclosing it. To this end, we compile SQL queries to be executed on client-side datasets, and automatically produce protocols that guarantee both integrity for the query results and confidentiality for the rest of the data. Our protocols rely on zero- knowledge cryptographic evidence, so that query evaluations can be checked without leaking information. We have built prototype compilers that produce C and F#; the F# code can be verified on the fly using our latest type systems for security. We illustrate our approach on queries for paying utility bills and pay-as-you-go insurance policies based on the detailed readings provided by smart meters and GPS units.

3.10 All-but-k Commitments

Ian Goldberg (University of Waterloo, CA)

Joint work of Henry, Ryan; Goldberg, Ian; Kate, Aniket; Zaverucha, Gregory; Olumofin, Femi; Huang, Yizhou

In this talk, we present a new kind of commitment scheme called "all-but- k commitments". With these commitments, Alice can commit to $n - k$ items, and then later open the commitment to n items. Bob is assured that Alice knew $n - k$ of those n items at the time of the commitment, but does not know which. We demonstrate an efficient implementation of all-but-k commitments using Kate et al.'s polynomial commitments from Asiacrypt 2010.

We illustrate the need for all-but-k commitments by demonstrating an attack on the soundness of Peng and Bao's "batch zero-knowledge proof and verification" protocol for proving knowledge and equality of one-out-of- n pairs of discrete logarithms. We repair the protocol using our new commitment construction, and in fact can easily generalize the repaired protocol to a k-out-of-n setting. For $k = 1$, this yields an "OR" proof; for $k = n$, this yields an "AND" proof. For intermediate values of k, this batch protocol is entirely novel.

3.11 Tightly Secure Signatures and Public-Key Encryption

Dennis Hofheinz (KIT – Karlsruhe Institute of Technology, DE)

Joint work of Hofheinz, Dennis; Jager, Tibor
Main reference D. Hofheinz, T. Jager, "Tightly Secure Signatures and Public-Key Encryption," in Advances in Cryptology (CRYPTO'12), LNCS, vol. 7417, pp. 590–607, Springer, 2011.
URL http://dx.doi.org/10.1007/978-3-642-32009-5_35

We construct the first public-key encryption scheme whose chosen- ciphertext (i.e., IND-CCA) security can be proved under a standard assumption and does not degrade in either the number of users or the number of ciphertexts. In particular, our scheme can be safely deployed in unknown settings in which no a-priori bound on the number of encryptions and/or users is known.

As a central technical building block, we devise the first structure-preserving signature scheme with a tight security reduction. (This signature scheme may be of independent interest.) Combining this scheme with Groth-Sahai proofs yields a tightly simulation-sound non-interactive zero-knowledge proof system for group equations. If we use this proof system in the Naor- Yung double encryption scheme, we obtain a tightly IND-CCA secure public-key encryption scheme from the Decision Linear assumption.

We point out that our techniques are not specific to public-key encryption security. Rather, we view our signature scheme and proof system as general building blocks that can help to achieve a tight security reduction.

3.12 On Genomic Privacy

Jean Pierre Hubaux (EPFL – Lausanne, CH)

Joint work of Ayday, E.; Fellay, J.; Hubaux, J.-P.; Humbert, M.; McAllen, P.; Raisaro, J. L.; Rougemont, J.

Over the last 20 years, DNA sequencing has progressed faster than Moore's law. This amazing progress is paving the way to so-called "personalized medicine", meaning that treatments can be fine-tuned to the genetic profile of each patient. In some hospitals, systematic sequencing of (consenting) patients is expected to happen as early as next year.

Yet, this evolution has dramatic implications in terms of privacy. If they happened to fall in the wrong hands, genomic data could be used to perpetrate various kinds of social discrimination (e.g. for decisions related to health/life insurance, recruitment, mortgage,...). Such data can also be used for paternity tests.

In this talk, we will provide an introduction to genomics for computer scientists and briefly describe the (few) ongoing research efforts in this field. We will also explain our own research project, focused on disease- susceptibility tests. Homomorphic encryption makes it possible to preserve diagnostic accuracy while protecting data privacy, at a reasonable cost.

Finally, we will show that the research challenges are formidable and that much more work needs to be done to meet the privacy challenges raised by genomics and more generally by the expected fundamental transformation of medicine [1].

References

1 E. Topol. *The Creative Destruction of Medicine*. Basic Books, 2012

3.13 LIRA: Lightweight Incentivized Routing for Anonymity

Aaron Johnson (Naval Research – Washington, US)

Joint work of Jansen, Rob; Johnson, Aaron; Syverson, Paul

Tor, the most popular deployed distributed onion routing network, suffers from performance and scalability problems stemming from a lack of incentives for volunteers to contribute. Insufficient capacity limits scalability and harms the anonymity of its users.

We introduce LIRA, a lightweight scheme that creates performance incentives for users to contribute bandwidth resources to the Tor network. LIRA uses a novel cryptographic lottery: winners may be guessed with tunable probability by any user or bought in exchange for resource contributions. The traffic of those winning the lottery is prioritized through Tor. The uncertainty of whether a buyer or a guesser is getting priority improves the anonymity of those purchasing winners, while the performance incentives encourage contribution. LIRA is more lightweight than prior reward schemes that pay for service and provides better anonymity than schemes that simply give priority to traffic originating from fast relays.

We analyze LIRA's efficiency, anonymity, and incentives, present a prototype implementation, and describe experiments that show it indeed improves performance for those servicing the network.

3.14 Towards Automizing Secure Two-Party Computation

Stefan Katzenbeisser (TU Darmstadt, DE)

The practical application of Secure Two-Party Computation is hindered by the difficulty to implement secure computation protocols. While recent work has proposed very simple programming languages which can be used to specify secure computations, it is still difficult for practitioners to use them, and cumbersome to translate existing source code into this format. Similarly, the manual construction of two-party computation protocols, in particular ones based on the approach of garbled circuits, is labor intensive and error-prone.

The talk describes the design of a new tool called CBMC-GC, which achieves Secure Two-Party Computation for ANSI C programs. Our work is based on a combination of model checking techniques and two-party computation based on garbled circuits. Our key insight is a nonstandard use of the bit- precise model checker CBMC which enables us to translate ANSI C programs into equivalent Boolean circuits. To this end, we modify the standard CBMC translation from programs into Boolean formulas whose variables correspond to the memory bits manipulated by the program. We demonstrate that CBMC-GC can compile reasonably-sized programs and achieves practical performance.

3.15 Ideal APIs and Modular Verification for TLS 1.2

Markulf Kohlweiss (Microsoft Research UK – Cambridge, GB)

TLS is possibly the most used secure communications protocol, and also the most studied, with a 18-year history of flaws and fixes, ranging from its protocol logic to its cryptographic design, and from the Internet standards to its numerous implementations. We develop a verified reference implementation of TLS 1.2. Our code fully supports its wire formats, ciphersuites, sessions and connections, re-handshakes and resumptions, alerts and errors, and data fragmentation, as prescribed in the RFCs; it interoperates with mainstream web browsers and servers. At the same time, our code is carefully structured to enable its modular, automated verification, from its main API down to computational assumptions on its sub-protocols and underlying cryptographic algorithms.

Our implementation is written in F# and specified in F7. We present its main interfaces—which may be read as specifications of large ideal functionalities—for its main components, such as authenticated encryption for the record layer and key establishment for the handshake. We describe its verification using the F7 refinement typechecker. To this end, we equip each cryptographic and construction of TLS with a new, ideal typed interface that captures its security properties, and we show how to replace their implementations with ideal functionalities while preserving indistinguishability. We thus obtain precise security results for a large part of TLS, for every byte of every connection, relative to the strength of the handshake and the record layer cryptography as selected by the negotiated ciphersuite. We also report new attacks.

3.16 On The Security of One-Witness Blind Signature Schemes, and on Some Alternatives

Anna Lysyanskaya (Brown University – Providence, US)

Joint work of Baldimtsi, Foteini; Lysyanskaya, Anna

Blind signatures have proved an essential building block for applications that protect privacy while ensuring unforgeability, e.g., electronic cash and electronic voting. One of the oldest, and most efficient blind signature schemes is the one due to Schnorr that is based on his famous identification scheme. Although it was proposed over twenty years ago, its unforgeability remains an open problem, even in the random-oracle model. In this talk, I will first show that current techniques for proving security in the random oracle model do not work for the Schnorr blind signature. Our results generalize to other important blind signatures, such as the one due to Brands. Brands' blind signature is at the heart of Microsoft's UProve system, which makes this work relevant to cryptographic practice as well.

This negative result naturally leads to the next question: Can we achieve the attractive features of UProve (lightweight, albeit linkable, anonymous credentials with attributes from just a couple of exponentiations in elliptic curve groups) in a provably secure fashion? Blind signatures alone do not give us the desired functionality; instead, we define blind signatures *with attributes* and give a construction for those whose efficiency is comparable to that of

UProve, and whose security relies on the decisional Diffie-Hellman assumption.

This talk is based on two joint papers with Foteini Baldimtsi:
– http://eprint.iacr.org/2012/197 and
– http://eprint.iacr.org/2012/298.

3.17 Relationships among Privacy Notions for Signatures

Mark Manulis (University of Surrey, GB)

Joint work of Fleischhacker, Nils; Günther, Felix; Kiefer, Franziskus; Manulis, Mark; Poettering, Bertram
Main reference N. Fleischhacker, F. Günther, F. Kiefer, M. Manulis, B. Poettering, "Pseudorandom Signatures," Cryptology ePrint Archive: Report 2011/673.
URL http://eprint.iacr.org/2011/673

Research on privacy for (ordinary) digital signatures, initiated by Yang et al. (PKC 2006) and continued by Fischlin (PKC 2007) demonstrated that for high-entropy hidden messages digital signatures can provide signer's anonymity. Later, Dent et al. (PKC 2010) showed that in this setting digital signatures can also provide confidentiality for the signed messages.

Building on these results I'll show that in fact digital signatures admit much stronger privacy guarantees than previously thought. It is namely possible to hide the entire information about the signature scheme, including its parameters, specification of algorithms, etc.

I'll talk about the new notion and constructions of *pseudorandom signatures*, which essentially guarantee that no adversary can distinguish between a string that was output by the signing algorithm and some randomly chosen bit string (of appropriate length). I'll demonstrate how this notion relates to existing notions of anonymity and confidentiality and propose efficient techniques that can be used to construct pseudorandom signatures.

This talk is based on a joint paper with Nils Fleischhacker, Felix Günther, Franziskus Kiefer, and Bertram Poettering, http://eprint.iacr.org/2011/673.

3.18 Practical Yet Universally Composable Two-Server Password-Authenticated Secret Sharing

Gregory Neven (IBM Research – Zürich, CH)

Joint work of Camenisch, Jan; Lysyanskaya, Anna; Neven, Gregory

Password-authenticated secret sharing (PASS) schemes, first introduced by Bagherzandi et al. at CCS 2011, allow users to distribute data among several servers so that the data can be recovered using a single human- memorizable password, but no single server (or even no collusion of servers up to a certain size) can mount an off-line dictionary attack on the password or learn anything about the data.

We propose a new, universally composable (UC) security definition for the two-server case (2PASS) in the public-key setting that addresses a number of relevant limitations of the previous, non-UC definition. For example, our definition makes no prior assumptions on the distribution of passwords, preserves security when honest users mistype their passwords, and guarantees secure composition with other protocols in spite of the unavoidable non-negligible

success rate of online dictionary attacks. We further present a concrete 2PASS protocol and prove that it meets our definition. Given the strong security guarantees, our protocol is surprisingly efficient: in its most efficient instantiation under the DDH assumption in the random-oracle model, it requires fewer than twenty elliptic-curve exponentiations on the user's device. We achieve our results by careful protocol design and by exclusively focusing on the two-server public-key setting.

3.19 Identifying Common Friends in a Privacy-Preserving Way

Bertram Poettering (RHUL – London, GB)

Joint work of De Cristofaro, Emiliano; Manulis, Mark; Poettering, Bertram
Main reference E. De Cristofaro, M. Manulis, B. Poettering, "Private Discovery of Common Social Contacts," Cryptology ePrint Archive: Report 2011/026
URL http://eprint.iacr.org/2011/026

The past decade witnessed a plethora of novel platforms and techniques for online social interaction, including different forms of online social networks, ubiquitous computing on smartphones, and so on. Clearly these developments also pose novel privacy threats on participants and their data.

This talk focuses on a specific problem that arises when users of a social network want to learn the set of their common friends in a privacy- preserving way, i.e. without disclosing non-matching contacts to each other, and without relying on a trusted third party (which might not be reachable in a ubiquitous environment). We offer a full cryptographic treatment of the problem, including security models and provably-secure solutions.

3.20 Data Privacy at Scale

Bartosz Przydatek (Google Switzerland – Zürich, CH)

As users entrust more and more private data to the cloud, it is critical to provide adequate protections for the data, yet without sacrificing high performance, availability and functionality of services, and without impeding innovation.

I will talk about challenges in protecting data privacy at scale, and will discuss a suite of technologies we developed to help addressing these challenges. In particular, I will touch upon issues caused by the lack of key management infrastructure accessible for the general public, and explain why even an efficient fully homomorphic encryption would probably not solve all the problems. I will describe then a simple scalable architecture in which data encryption with appropriate authorization checks and logging provides a meaningful protection for data privacy.

3.21 New Privacy Issues in UMTS

Jean-Pierre Seifert (TU Berlin, DE)

Joint work of Arapinis, Myrto; Mancini, Loretta; Ritter, Eike; Ryan, Mark; Golde, Nico; Redon, Kevin; Borgaonkar, Ravishankar

Mobile telephony equipment is daily carried by billions of subscribers everywhere they go. Avoiding linkability of subscribers by third parties, and protecting the privacy of those subscribers is one of the goals of mobile telecommunication protocols.

We use formal methods to model and analyse the security properties of 3G protocols. We expose two novel threats to the user privacy in 3G telephony systems, which make it possible to trace and identify mobile telephony subscribers, and we demonstrate the feasibility of a low cost implementation of these attacks. We propose fixes to these privacy issues, which also take into account and solve other privacy attacks known from the literature. We successfully prove that our privacy-friendly fixes satisfy the desired unlinkability and anonymity properties using the automatic verification tool `ProVerif`.

3.22 SSL: Myth or Reality?

Vitaly Shmatikov (University of Texas at Austin, US)

Originally deployed in Web browsers, SSL (Secure Sockets Layer) has become the de facto standard for secure Internet communications and is now used widely even in non-browser software. SSL is intended to provide end-to-end security even against an active, man-in-the-middle attacker.

It turns out that SSL is completely insecure against a man-in-the-middle attack in many critical applications and libraries. Vulnerable software includes Amazon's EC2 Java library and all cloud clients based on it, Amazon's and PayPal's merchant SDKs responsible for transmitting payment details from e-commerce sites to payment gateways, integrated shopping carts such as osCommerce, ZenCart, Ubercart, and PrestaShop, AdMob code used by mobile websites, Chase mobile banking and many other Android apps and libraries, as well as Java Web-services middleware and all applications based on it. Interestingly, all these programs use correct SSL implementations... badly.

I will discuss the root causes of these vulnerabilities and present some recommendations for the developers of SSL libraries and applications that use SSL.

This talk is based on the forthcoming ACM CCS 2012 paper.

3.23 Practical Oblivious Access at 1Mbps+

Radu Sion (Stony Brook University, US)

We review several of our recent results including the first/fastest practical ORAM linux file system as well as other fun cloud-related crypto. This is also a preview of our CCS 2012 ORAM papers.

3.24 Anonymity and One-Way Authentication in Key Exchange Protocols

Douglas Stebila (Queensland University of Technology, AU)

Main reference I. Goldberg, D. Stebila, B. Ustaoglu, "Anonymity and one-way authentication in key exchange protocols," in *Designs, Codes and Cryptography*.
URL http://dx.doi.org/10.1007/s10623-011-9604-z

Key establishment is a crucial cryptographic primitive for building secure communication channels between two parties in a network. It has been studied extensively in theory and widely deployed in practice. In the research literature a typical protocol in the public-key setting aims for key secrecy and mutual authentication. However, there are many important practical scenarios where mutual authentication is undesirable, such as in anonymity networks like Tor, or is difficult to achieve due to insufficient public-key infrastructure at the user level, as is the case on the Internet today.

In this work we are concerned with the scenario where two parties establish a private shared session key, but only one party authenticates to the other; in fact, the unauthenticated party may wish to have strong anonymity guarantees. We present a desirable set of security, authentication, and anonymity goals for this setting and develop a model which captures these properties. Our approach allows for clients to choose among different levels of authentication. We also describe an attack on a previous protocol of Overlier and Syverson, and present a new, efficient key exchange protocol that provides one-way authentication and anonymity.

3.25 Privacy-Preserving Interval Operations

Susanne Wetzel (Stevens Institute of Technology, US)

Joint work of Mayer, Daniel; Meyer, Ulrike; Wetzel, Susanne

In this talk we present some work-in-progress on designing privacy- preserving protocols for operations on intervals of integers. In particular, we will present new 2-party protocols to test for the overlap of two intervals, to determine the size of the overlap of two intervals, and to select a random sub-interval in the overlap. We will show that the new protocols are privacy- preserving in the context of a semi-honest adversary.

3.26 Recent Attacks On Mix-Nets

Douglas Wikstroem (KTH Stockholm, SE)

We revisit mix-nets with randomized partial checking (RPC) as proposed by Jakobsson, Juels, and Rivest (2002). RPC is a technique to verify the correctness of an execution both for Chaumian and homomorphic mix-nets. We identify serious issues in the original description of mix-nets with RPC and show how to exploit these to break both correctness and privacy. Our attacks are practical and applicable to real world mix-net implementations, e.g., the Civitas and the Scantegrity voting systems.

We also consider the heuristically secure mix-net proposed by Puiggalí and Guasch (EVOTE 2010) used in the recent large scale electronic elections in Norway. We present practical attacks on both correctness and privacy for some sets of parameters of the scheme. Currently, we are unable to leverage this into an attack on the electronic election scheme as a whole due to additional components.

4 Working Groups

Many participants took the opportunity offered by Dagstuhl's cosy atmosphere to work face-to-face in smaller groups. This work included continuation of existing research collaborations, engagement in smaller group discussions, and initiation of new research agendas.

5 Open Problems

One of the main open problems, as considered by many seminar participants, is that existing privacy models are often too narrow and do not address various privacy threats that exist in the real world. As a result, development of more sophisticated privacy models and appropriate design of provably private yet practically relevant privacy-oriented security protocols and mechanisms was identified as an important research direction for the coming years.

6 Panel Discussions

The seminar included two panel discussions on the models and definitions of privacy and on the use of privacy-oriented cryptography in practice. These panels discussion are summarized in the following.

6.1 Panel "Privacy Models: UC or Not UC?"

Moderator: Jan Camenisch
Panelists: Giuseppe Ateniese, Yevgeniy Dodis, Ian Goldberg, Dennis Hofheinz

There are a number of different approaches for proving the security of a cryptographic protocol or primitive. The probably most popular ones are game based definitions and the universal composability (UC) model. In the former model, security is defined by a number of games, each capturing one (security) property that the protocol/primitive should satisfy.

This approach typically leads to shorter and easier to read proofs but has the drawback the protocol/primitive might satisfy each property separately but not all of the at the same time. In contrast, in the UC model, security is defined by specifying an ideal process describing the expected behavior of the protocols and then it is proved that the participants of the protocol cannot distinguish whether they interact with the ideal process or the protocol.

The panelist and the audience vividly discusses the topic, not only during the panel but also throughout the whole workshop. There seems to have been a common agreement that specifying a protocol by an ideal process is the right thing to do. It captures precisely what a protocol achieves on the one hand and in principle provides users of cryptography such as system designers with an accessible description of what a protocol achieves and how is can be used. There also seemed to have been agreement that the current UC-style models do not quite achieve this yet for a number of reasons. First, the specifications are often very complex and therefore hard to understand so that still it's not easy for systems designer to employ cryptography. Second, the proofs in the UC-style models are often much much longer that in game based definitions. Thus, the proofs are hard to write and hard to verify (probably very few people apart from the authors read them). The panelists and audience felt that it would be nice if:

- Guidelines for defining an ideal process existed, similar to, e.g., JAVA programming language textbooks.
- That probably a second abstraction layer (or at least explanation) is needed that makes a ideal process specification more accessible, together with a (provable) reduction to the ideal process specification.
- Better tools to structure proofs existed so that they are easier to write and read and therefore more confidence in the correctness of the proofs is achieved.

6.2 Panel "Privacy-Oriented Cryptography: Why is it not adopted more in practice?"

Moderator: Rebecca N. Wright
Panelists: Emiliano de Cristofaro, George Danezis, Anna Lysyankaya, Kai Rannenberg, and Radu Sion
Invited Commenters: Roger Dingledine and Ian Goldberg

Privacy-oriented cryptography can provide enhanced privacy protection for individuals, businesses, and governments in a large variety of situations, including web personalization, smart metering, health care, and surveillance. Beyond just protecting the privacy of information in transit or in storage, privacy-oriented cryptography enables "computing without revealing", in which parties can carry out certain computations or interactions, while only learning specific results. This panel addressed the extent to which privacy-oriented cryptography is and is not adopted for practical use, and barriers that must be overcome in order for it to have increased adoption.

The panelists and other participants noted that in fact there has already been some adoption of privacy-oriented cryptography in practice, citing, for example: Open SSL, which has massive deployment and use; German identity cards, which contain privacy-friendly cryptography; the ISO definition of identity, which is based on attributes only; and the Tor project, which provides anonymous Internet communication. Additionally, some felt that it is simply too early in the development of such technologies and supporting technologies

on which they rely (e.g., fast communications and large storage capabilities) to expect widespread adoption. As a comparison, nuclear energy has been around for decades, and it still faces technology problems and resistance. Evidence toward continued progress are that companies are beginning to pay attention to the value of adopting privacy technologies, (e.g. MSIE and DoNotTrack, encryption in the cloud and in telcos), and governments have programs (NIST & Trust identities, DARPA & computation over encrypted data) developing and standardizing these technologies.

The panelists and other participants also discussed factors that potentially limit the practical adoption of privacy-oriented cryptography and how they might be overcome. These include the following:

- Privacy-oriented cryptography solutions are difficult to understand, and cryptographers don't spend enough time explaining their ideas. Some noted that the job of cryptographers is to push the state-of-the-art of designs, and let professional engineers incorporate these into systems in ways that cryptographers can't necessarily predict. Others noted that there is a credibility gap between cryptographers and engineers. To bridge this gap, engineers need to understand the assumptions and tools of cryptography, and cryptographers need to take engineering concerns seriously. This takes willingness, time, and effort by both cryptographers and engineers, and may take decades.
- Researchers do not typically carry the results far enough into the technology pipeline to achieve real-world deployment. This is partly caused by funding and training incentives for researchers to focus on publishing papers and moving on to new results rather than carrying individual ideas through to deployable and deployed systems. Further, it is sometimes easier to publish papers that break designs than that build them. As a result, privacy technology designs often do not consider how end users will understand and interact with them, further limiting their potential for adoption.
- There is a real gap between problems that theoretically seem like privacy-preserving cryptography can solve and the actual realities that are revealed when adoption is attempted. This requires feedback with the community that must respond to new understandings of the challenges to be solved. Even when developed, privacy technology is often difficult to understand and use, even for experts. This means even when the technology is used, it is often used incorrectly.
- People always say they value privacy, but quickly lose interest when they must pay for it, learn how to use new technologies, or otherwise change their behavior.
- Companies are not incentivized to protect consumer privacy, only business privacy. There is a perception (often a misperception, we think) that consumer data is a gold mine and privacy technology and legislation can only hurt profits. Businesses only protect privacy to the extent that their (largely under-informed) customers demand it.
- Privacy legislation may be the best way to drive large-scale deployment of privacy technologies. However, convincing governments to adopt technology or legislation is a political process. This is not something that researchers are typically well-equipped to handle, both by training and by funding.
- Free technology is more likely to be adopted than technologies that individuals or companies must pay for, but these must be paid for somehow. Example successes paid for by government funding and volunteer time of an implementation/adoption community, motivated by keeping users safe: Open SSL, Off-the-Record messaging, Tor.

Participants

- Giuseppe Ateniese
Johns Hopkins University – Baltimore, US
- Johannes Blömer
Universität Paderborn, DE
- Nikita Borisov
Univ. of Illinois – Urbana, US
- Xavier Boyen
Prime Cryptography – Palo Alto, US
- Jan Camenisch
IBM Research – Zürich, CH
- Claude Castelluccia
INRIA Rhône-Alpes, FR
- Bruno Crispo
University of Trento – Povo, IT
- George Danezis
Microsoft Research UK – Cambridge, GB
- Emiliano De Cristofaro
PARC – Palo Alto, US
- Claudia Diaz
K.U. Leuven, BE
- Roger Dingledine
The Tor Project, US
- Yevgeniy Dodis
New York University, US
- Maria Dubovitskaya
IBM Research – Zürich, CH
- Marc Fischlin
TU Darmstadt, DE
- Cédric Fournet
Microsoft Research UK – Cambridge, GB
- Ian Goldberg
University of Waterloo, CA
- Dennis Hofheinz
KIT – Karlsruhe Institute of Technology, DE
- Jean Pierre Hubaux
EPFL – Lausanne, CH
- Aaron M. Johnson
Naval Res. – Washington, US
- Stefan Katzenbeisser
TU Darmstadt, DE
- Dogan Kesdogan
Universität Siegen, DE
- Markulf Kohlweiss
Microsoft Research UK – Cambridge, GB
- Anja Lehmann
IBM Research – Zürich, CH
- Anna Lysyanskaya
Brown Univ. – Providence, US
- Mark Manulis
University of Surrey, GB
- Gregory Neven
IBM Research – Zürich, CH
- Bertram Poettering
RHUL – London, GB
- Bartosz Przydatek
Google Switzerland – Zürich, CH
- Kai Rannenberg
Goethe-Universität Frankfurt am Main, DE
- Jean-Pierre Seifert
TU Berlin, DE
- Vitaly Shmatikov
University of Texas at Austin, US
- Radu Sion
Stony Brook University, US
- Douglas Stebila
University of Technology Brisbane, AU
- Gene Tsudik
Univ. of California – Irvine, US
- Markus Ullmann
BSI – Bonn, DE
- Susanne Wetzel
Stevens Inst. of Technology, US
- Douglas Wikström
KTH Stockholm, SE
- Rebecca Wright
Rutgers Univ. – Piscataway, US

Report from Dagstuhl Perspectives Workshop 12382

Computation and Palaeography: Potentials and Limits

Edited by

Tal Hassner[1], Malte Rehbein[2], Peter A. Stokes[3], and Lior Wolf[4]

1 **Open University – Israel, IL,** `hassner@openu.ac.il`
2 **Universität Würzburg, DE,** `malte.rehbein@uni-wuerzburg.de` **/ University of Nebraska Lincoln, US,** `malte.rehbein@unl.edu`
3 **King's College London, GB,** `peter.stokes@kcl.ac.uk`
4 **Tel Aviv University, IL,** `wolf@cs.tau.ac.il`

Abstract

This report documents the program and outcomes of Dagstuhl Seminar 12382 'Perspectives Workshop: Computation and Palaeography: Potentials and Limits'. The workshop focused on the interaction of palaeography, the study of ancient and medieval documents, with computerized tools, particularly those developed for analysis of digital images and text mining. The goal of this marriage of disciplines is to provide efficient solutions to time-consuming and laborious palaeographic tasks. It furthermore attempts to provide scholars with quantitative evidence to palaeographical arguments, consequently facilitating a better understanding of our cultural heritage through the unique perspective of ancient and medieval documents. The workshop provided a vital opportunity for palaeographers to interact and discuss the potential of digital methods with computer scientists specializing in machine vision and statistical data analysis. This was essential not only in suggesting new directions and ideas for improving palaeographic research, but also in identifying questions which scholars working individually, in their respective fields, would not have asked without directly communicating with colleagues from outside their research community.

Seminar 18.–21. September, 2012 – www.dagstuhl.de/12382
1998 ACM Subject Classification I.7.5 Document Capture, J.5 Arts and Humanities
Keywords and phrases Digital Palaeography, Document Analysis
Digital Object Identifier 10.4230/DagRep.2.9.184

1 Executive Summary

Tal Hassner
Malte Rehbein
Peter A. Stokes
Lior Wolf

The Schloss-Dagstuhl Perspectives Workshop on "Computation and Palaeography: Potentials and Limits" focused on the interaction of palaeography, the study of ancient and medieval documents, and computerized tools developed for analysis of digital images in computer vision. During the workshop, the interaction between domain experts from palaeography and computer scientists with computer vision backgrounds has yielded several very clear themes for the future of computerized tools in palaeographic research. Namely,

Computation and Palaeography: Potentials and Limits, *Dagstuhl Reports*, Vol. 2, Issue 9, pp. 184–199
Editors: Tal Hassner, Malte Rehbein, Peter A. Stokes, and Lior Wolf
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

- Difficulties in communication between palaeographers and computer scientists is a prevailing problem. This is often reflected not only in computerized tools failing to meet the requirements of palaeography practitioners but also in the terminology used by the two disciplines. Better communication should be fostered by joint events and long-term collaborations.
- Computerized palaeographic tools are often black boxes which put the palaeography scholar on one end of the system, only receiving a systems output, with little opportunity to directly influence how the system performs or to communicate with it using natural palaeographic terminology. The long-term desire is to have the scholar at the center of the computerized system, allowing interaction and feedback in order to both fine-tune performance and better interpret and communicate results. This is crucial if palaeography is to become a truly evidence-based discipline. To this end the use of high-level terminology, natural to palaeography, should be integrated into computerized palaeographic systems.
- Palaeographic data, scarce to begin with, is even more restricted by accessibility and indexing problems, non-standard benchmarking techniques and the lack of accurate meta-data and ground truth information. Multiple opportunities were identified for acquiring data and disseminating it both in the palaeographic research community and outside to the general public.
- Palaeographic research is largely restricted to the domain of experts. Making palaeography accessible to non-experts by using computerized tools has been identified as an effective means of disseminating valuable cultural heritage information while at the same time potentially giving rise to crowdsourcing opportunities, such as those proved successful in other domains.

The manifesto which resulted from this work elaborates on the existing challenges and limitations of the field and details the long-term recommendations that have emerged from the workshop.

2 Table of Contents

3 Overview of Talks

3.1 Three Pattern-Recognition Approaches to the Automatic Identification of the Writers of Ancient Documents

Dimitris Arabadjis and Micalis Panagopoulos (National TU – Athens, GR)

Joint work of Papaodysseus, Constantinl; Arabadjis, Dimitris; Rousopoulos, Panayiotis; Giannopoulos, Fotios; Blackwell, Chris

Dating the content of ancient documents is absolutely crucial for History and Archaeology. For example one of the most prominent historians, Professor Christian Habicht, has recently written that proper historical use of inscriptions can only be made if they can be dated. However, writers of ancient inscriptions and manuscripts, as a rule, did not sign or date their documents. So far, dating the content of ancient inscriptions and manuscripts is a very difficult task and is based on scholars' instinct and frequently subjective considerations. One main goal of the work that was presented is to perform quantitative analysis on the scribal hands, so that the relationships among these volumes and their relative dates of production are obtained. This will be achieved by means of writer identification, since the working careers of most ancient writers covered about 20 to 25 years. So, if one could attribute a document to a writer, then the content of the document gains a date immediately, which is clearly the time period during which the writer was active. Hence, three different approaches for the identification of the writer of ancient documents were outlined.

The first approach estimates ideal representatives of selected alphabet symbols for each document separately and then compares these representatives. The second approach introduces a new mathematical notion, a kind of two dimensional curvature, so that the various alphabet symbols realizations can be optimally matched and compared after proper associated transformations. The third approach uses exhaustive comparisons based on classical curvature and set-theoretic similarity measures. In addition, a number of cases were presented where identification of the writer(s) of the ancient documents is of great importance for the analysis of their content.

3.2 Multi-Source and Multi-View 3D Data Exploration

Matthieu Exbrayat (Université d'Orleans, FR)

Joint work of Exbrayat, Matthieu; Martin, Lionel

Multi-source data consists of data, coming from various producers, that share a common object of interest. During the 2008-2011 Graphem Project, we have been studying the spatial visualisation of medieval writing samples, by the means of an interactive spatial projection tool named Explorer3D that we developed at LIFO. In this tool we take as an input the description of the writing samples, in the form of a set of numerical features. Based on these features, we use 3D projection techniques, such as Principal Component Analysis, to create a 3D space in which each sample is represented by a point, so that the distance between points in the 3D space reflects the proximity (or distance) of the writing samples according to the input features. We offer various interaction extensions, for instance to visualise the writing

samples or to modify the 3D projection based on the user's visual analysis of the relevance of this projection.

The feature sets we based our projections on during Graphem were produced by the other computer science partners of this project. Several sets of features have been proposed and evaluated. Nevertheless each of them has been studied independently. Arguing that a visual comparative study of these feature sets might help understanding their respective strengths and weaknesses, we have extended Explorer3D in order to load and visualise a set of writing samples using several feature sets simultaneously, and thus using several 3D scenes. Various interactive tools have been developed or adapted in order to compare how writing samples appear similar or dissimilar across the available 3D scenes, and thus across the underlying feature sets.

3.3 Computerized Paleography of Hebrew Writing from the First Temple Period

Shira Faigenbaum (Tel Aviv University, IL)

The only texts from the First Temple Period in Israel and Judah that endured the harsh local climate were written in ink on pieces of pottery (ostraca). The discipline of Iron Age epigraphy classically involves manual labor in analyzing the inscriptions and establishing a comparative typology of characters. However, this approach may unintentionally mix documentation and interpretation. We introduce image processing and pattern recognition methods to the field of First Temple epigraphy, minimizing the epigrapher's involvement in activities prone to subjective judgment. Our work comprises various image acquisition techniques, image quality assessment, image binarization, and letter comparison metrics.

3.4 Modern Technologies for Manuscript Research

Melanie Gau and Robert Slabatnig (TU Wien, AT)

This paper on multispectral imaging and image enhancement addressed the following:

- Making faded-out text legible
- Unmixing and inpainting techniques for palimpsests
- Stroke and Character Analysis
- Information about writing tools
- Degraded Character Recognition and OCR
- Layout Analysis and Text Line Detection

3.5 Experiments in the Digital Humanities

R. Manmatha (University of Massachusets – Amherst, US)

Joint work of Manmatha, R.; Rath, Toni; Rothfeder, Jamie; Lavrenko, Victor; Yalniz, I. Zeki; Can, Ethem
Main reference I. Yalniz, E. Can, R. Manmatha, "Partial duplicate detection for large book collections," in Proc. of the 20th Conf. on Information Knowledge and Management (CIKM'11), pp. 469–474, ACM, 2011.
URL http://doi.acm.org/10.1145/2063576.2063647
URL http://ciir.cs.umass.edu/irdemo/hw-demo/

Searching historical handwritten manuscripts is a challenging task given that there is a way to go before handwriting recognition is reasonably accurate. I described two approaches to doing this. The first is word spotting – where the query is a word image – and the system looks for similar word images in a set of documents. Word spotting has now been investigated by a large number of researchers for handwritten and printed documents. The second is based on relevance models where the word images are automatically annotated with vocabulary words and their probabilities using a joint probability model. I showed a demonstration of such an automatic system on a sample of George Washington's documents. Such an automatic system requires a number of other automatic steps including automatic word segmentation and I described a technique for doing this.

In the second part of the talk I described an efficient automatic approach to linking old printed (and scanned) books by finding partial duplicates. For example, in a large collection of books one can find different versions of Shakespeare's Othello or Virgil's Aeneid. One version may just have the main text, a second scholarly version may have a lot of footnotes while a third may have substantial additional text in the form of an introduction and endnotes. By representing a book as a sequence of unique words one can find partial duplicates efficiently. A similar approach may be used to find translations of books without explicitly translating a book.

3.6 Challenges in Palaeography for which Computer Sciences Might Offer Some Solutions

Wendy Scase (University of Birmingham, GB)

The aim of my presentation was to facilitate identification of kinds of important palaeographical problem that computer sciences might make a major contribution to solving in the next few years. I illustrated a variety of problems that I have encountered in my own projects on Middle English manuscripts and offered a rudimentary analysis of some of the different kinds of problem palaeographers and manuscripts specialists are faced with. The problems discussed were of three kinds.

1) Problems in the analysis of digital images of medieval manuscripts that might be amenable to computer vision methods. Digitisation of manuscripts is proceeding fast and has transformed researchers' access to manuscripts but to maximise the benefit of digitised manuscript images for research, I proposed, we need to provide researchers with computer-assisted ways to search and analyse the images. I illustrated the method used for my Vernon manuscript project (Oxford, Bodleian Library, MS Eng.poet.a.1) where a full transcription and detailed manuscript description link to digital images of the entire manuscript greatly

assisting searching and analysis of this very large and important codex. I proposed that datasets such as this might provide training data for the development of computer-assisted recognition and searching so that such metadata could be created in a less labour-intensive way in future.

2) Problems in the linking of manuscript metadata and a possible solution. Much metadata about manuscripts is in digital form; often this is in discrete and distributed datasets. In recent years strides have been made to link this data (the example given was the Manuscripts Online project hosted at the University of Sheffield) opening up the possibility of conducting research across much larger datasets in future. However problems remain because different datasets use different vocabularies and languages. Tackling this by conventional means would be a huge task (as attempts have shown) and I suggested that it might be possible to semi-automate the making of dictionaries and thesauri of synonyms and related terms. I proposed it might be possible to harness the search activity of expert searchers to develop thesauri of synonyms and related terms to aid the search function. I proposed that perhaps the problem might be addressed by technology similar to that which underlies commercial search engines where systems harvest user activity and build up data on related and synonymous search terms that they then offer to users.

3) Problems in the Public Understanding of Manuscripts. I proposed that one of the most challenging and urgent problems is to improve public understanding and valuing of medieval manuscripts if they are to be a useful cultural resource in future and not to remain only available to a very small elite as they were when they were first made. I described successful forms of public engagement with manuscripts (particularly the Vernon MS) and asked for suggestions of how gamification and other computer-based strategies might be harnessed for this purpose.

3.7 Bringing the Digital to Palaeography: Some Background and Challenges

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

The purpose of this talk was to set the scene for the following days' discussion. It provided a brief overview of the recent history of the field, in an attempt to identify the status quo. There was no attempt to canvass all the projects to date, but instead to analyse the problems that they have addressed and the problems that still remain. In particular, it was argued that there is still a significant distance between 'computational' projects, lead principally by computer scientists, and 'palaeographical' projects lead by humanities scholars. Not only do they take different approaches, but the questions being asked by the two groups are often very different as well, and as a result neither group has had very much influence in the day-to-day research of the other. These difficulties are in addition to ones that have been previously identified in the literature, such as questions of trust, transparency and verifiability.

In order to start discussion a number of suggestions were made regarding:

- Which sorts of question are most amenable to computational methods from the position of a humanities scholar, and why?
- Which sorts of questions are interesting to humanities scholars but are not yet being addressed computationally (but could be)?

- What is the future of 'computational' vs other forms of 'digital' palaeography?
- Can related discussions in other branches of Digital Humanities help?

3.8 The Graphem Research Project

Dominique Stutzmann (CNRS – Paris, FR)

Joint work of Stutzmann, Dominique; Smith, Marc; Muzerelle, Denis; Gurrado, Maria; Eglin, Véronique; Bres, Stéphane; Lebourgeois, Frank; Joutel, Guillaume; Daher, Hani; Vincent, Joutel; Leydier, Yann; Exbrayat, Mathieu; Martin, Lionel; Moalla, Ikram; Siddiqi, Imran
Main reference D. Muzerelle, M. Gurrado, (eds), "Analyse d'image et paléographie systématique: travaux du programme 'Graphem'," Paris, Association 'Gazette du livre médiéval', 2011.

The research program GRAPHEM (Grapheme based Retrieval and Analysis for PalaeograpHic Expertise of medieval Manuscripts, 2008–2011), funded by the French National Research Agency, aimed at improving the data mining and image processing techniques applied to medieval scripts and their classification with several methods: outline directions, generalized cooccurrences, stroke categorization. The results of the unsupervised categorization showed too much overlapping to be used properly by academic end users from the humanities. The supervised methods reached very satisfying results in assigning a random handwriting to a category of script and to a century. Nevertheless the global approach from the methods of cooccurrences or curvelets makes it very difficult to relate the results back to the visual differences that the palaeographer can observe. Even in the chain code method (stroke categorisation), the calculated features cannot be traced back to the morphological ones that the palaeographers are used to observing. These features are observed through other methods: metrology (human-based measurements of writing features such as writing angle, density, word spacing), graphonomics, and (allo)graphetic transcriptions, which are more directly relevant to state the script evolution and also may have positive incidence on image analysis and text recognition.

3.9 The Ongoing Effort to Reconstruct the Cairo Genizah

Lior Wolf (Tel Aviv University, IL)

Many significant historical corpora contain leaves that are mixed up and no longer bound in their original state as multi-page documents. The reconstruction of old manuscripts from a mix of disjoint leaves can therefore be of paramount importance to historians and literary scholars. In collaboration with the The Friedberg Genizah Project, we showed that visual similarity provides meaningful pair-wise similarities between handwritten leaves and then went a step further to suggest a semi-automatic clustering tool that helps reconstruct the original documents. The proposed solution is based on a graphical model that makes inferences based on catalog information provided for each leaf as well as on the pairwise similarities of handwriting. Several novel active clustering techniques were explored, and the solution has been applied to a significant part of the Cairo Genizah, where the problem

of joining leaves remains unsolved even after a century of extensive study by hundreds of human scholars.

4 Working Groups

4.1 Acquisition of Images

Dimitris Arabadjis, Shira Faigenbaum, Robert Sablatnig, and Timothy Stinson

Standards for digital image acquisition need to be clearly articulated and the same protocol followed by all digital imaging projects when possible. These include practices such as:

1. Using color bars and grey cards
2. Documenting illumination used (e.g., how many lamps, their positioning, diffuser used)
3. Including references to size of original objects
4. Documenting information about photographic equipment used
5. Using shared standards for metadata descriptions of digitized objects
6. Including information that links multiple names and catalogue records when original objects have no single identifier (e.g., a manuscript with shelfmarks that change over time and that is also referred to by other common names in scholarly literature)
7. Establishing file naming conventions in order to facilitate the creation of good metadata and their proper sequence of images when books or other documents are being digitized.

Additionally, if one takes several images of the same object (e.g., jpeg, tiff, multiple sizes, multispectral), it is important that metadata indicates that these are images of the same object.

It would be helpful to have a set of guidelines articulating how to capture digital and analogue images across a wide range of technologies – e.g., scanning objects and photographic negatives, using digital and analogue cameras, digitizing microfilm.

Copyright or contractual use restrictions on photographs of cultural heritage items create many barriers for researchers. In many cases, tax-funded or state-supported research projects must expend significant financial and human resources on negotiating and paying for reproduction rights, even if those rights are being obtained from state repositories. Furthermore, rights tend to be granted only to scholars or research groups on a one-by-one basis, which frustrates large-scale studies of collections of manuscript images. Making large sets of images more easily available at an international scale would greatly facilitate the pursuit of significant new research questions (e.g., large-scale comparative studies of handwriting that map regional and national developments of hands across time).

It might be useful to call attention to libraries and museums with progressive policies that help researchers, such as the Austrian State Library, which makes images paid for by one project freely available to subsequent researchers needing those images.

4.2 Tools

Nachum Dershowitz, Matthieu Exbrayat, Eyal Ofek, Micalis Panagopoulos, and Ségolène Tarte

What tools are needed to progress in this field? The assets of computers are their ability to deal with big data, using memory, distinction/identification of fine differences, and rare occurrences. The assets of humans are dealing with complex data, making sense of the data, and gestalt questions. It is vital that these two sets of assets are combined through semi-automatic and interactive tools, not through 'black boxes': we must always keep humans in the loop! This includes

- Provide training data / annotated data
- Online training / expert-in-the-loop
- Crowd-sourcing

Rather than a single product, we also need a collection of tools that contribute to each other: a toolbox to account for the different needs of different researchers.

Low-level tools:

- Binarization – segmentation – alignment / matching / registration (for later comparison) – physical feature extraction – expert feature extraction (angles, curvatures, strokes...) – similarity measures (for comparison between characters, words, texts, fragments, documents, corpora)

Medium-level tools:

- Clustering – classification – character recognition – word spotting – searching (text via string – text via image – image via text – image via image – characters) – image-text correspondence
- Databases: organisation of data in a way that allows fast queries of metadata, transcripts, text qualities, etc.

Higher-level tools:

- Interfaces, ergonomy (CHI) – searches of combinations of characters/words (bigrams, trigrams) – correspondence of expert vocabularies – inferences of paraphrases and synonyms for searches through metadata
- A transcription tool to make the connection between text as shape and text as meaning

Other principles for development:

- Feedback loops and cognitive triggers: drawing/touch screen technologies – simple interactive image enhancements – visualization aspects of interactions with these tools (of results, of databases) – interactive visualisations (e.g. time varying graphs) – customizable visualisations – multiple languages – rationale building support, tracking of expert hypotheses in interpretation building – statistical tools with tests of significance – information sharing – sounding the texts
- Web-services to provide access to such tools via internet?

This topic has potential links with medical imaging, cognitive sciences, CHI, and NLP, all of which should be explored in future work.

4.3 Content and Context

Melanie Gau, R. Manmatha, Ophir Münz-Manor, Wendy Scase, and Dominique Stutzmann

Scholars need links from image to text (and vice versa): many manuscripts are already imaged but are not accessible in any way except to look at. Linking then becomes the key issue. This should ideally be done automatically and involving multiple forms of content, including not only images and transcripts but also other metadata such as contextual information, art references, articles and papers, content/semantics, codicology, textology, other discreet distributed datasets, named entities, descriptions, and so on.

The question, then, is how. A broad variety and combination of technical approaches and tools is required, e.g. word spotting, finding named entities (both using underlying dictionaries and also via more visionary approaches such as crawling the internet), spotting symbols, controlled vocabularies, text alignment between different versions of the same text such as old or even possibly faulty transcripts (also for acquiring training data), reference corpora, standardised datasets, handwriting recognition, alignment techniques, automatic creation of thesauruses to support queries/resource discovery. It should be possible to find all instances of a word in all images and texts, and to map vocabulary relations/keywords/concepts applicable to and mixable with different data sets, languages, collections, and intersections of information.

We would also like to recommend a note on an EU-wide harmonisation of copyright given the very wide range of policies and freedoms/restrictions across different institutions let alone countries.

4.4 Challenges and Limitations

Dimitris Arabadjis, Melanie Gau, and Ségolène Tarte

We face challenges, rather than limitations, in that the issues discussed here are not necessarily insurmountable. Technical limitations are not reviewed here because, in the light of the potential communications problems, they seem largely surmountable. The discussions and round table in this workshop has revealed that that a lot more is possible than single experts could predict, so any prognosis of technical limitations could have the risk of pre-emptive delimitations.

Current problems include computational limitations, access to data, issues of data retrieval, flexibility of searches (too flexible or too rigid – precision and recall). Major bottlenecks include communication, namely differences of terminology not only between disciplines, but also within disciplines due to different traditions in various specialities (eg. classics, slavonic studies, medieval studies all have different traditions). The tradition in computer science for image processing vs data mining has expert vocabularies (within a given field) which are a very abstract way of formulating problems that might not translate well into formal language. Mid-level features might be a useful compromise – a slowing down approach. This has the disadvantage of likely constraining the potential of each discipline, but better alternatives have not yet been found. A meeting ground is needed. Computer science has a

convention of not deriving natural interpretation from the methodology. What is excluded from systematic analysis at the moment is context and meaning, which are crucial (indeed, the whole point) for palaeographers. The output needs further cognitive processing to be interpreted, and computer science doesn't really have ways to do that, nor a tradition to do that. Instead there is need for systematisation and formatting of approaches: this will lead to better exchange but at the expense of less room for creativity both for palaeography and computer science. Nevertheless a common language is needed, for example for features in image processing vs features in palaeography. This in turn leads to issues of trust vs anxiety about black-boxes. Mutual education is also needed for this: an understanding on both sides of the main principles if not of detailed methodologies. For this, the middle-person/translator becomes vital. There is need either for an extra person for the role, or for one (at least) of the experts to be trained. This is the end of the age of the lone scholar in the humanities as well as in computer science.

4.5 Relevance to Society

Wendy Scase, Eyal Ofek, and Ophir Münz-Manor

Manuscripts are one of the major sources of knowledge of human culture and society for most of history. All of the world's written heritage produced before the invention of printing is handwritten. Much written heritage dating from after the introduction of printing is also in manuscript form. Unlike printed texts, manuscript sources are often highly inaccessible. They pose challenges of legibility, of interpretation, of language, and of subject matter. Owing to these challenges manuscript materials are often accessible for only a very small number of highly- trained expert groups. They also pose challenges of discovery and physical accessibility. There are hundreds of thousands of manuscript materials. They are scattered across the world in libraries, archives, museums, and private collections and no single catalogue or list exists to discover this material. Each manuscript source is unique and requires specialist curation and conservation. Exposing these materials to too much handling could result in damage and destruction.

For these reasons, despite their importance to knowledge of human history and culture, manuscripts have remained a largely untapped cultural resource. Exceptions to this neglect, such as the Book of Kells (now in Trinity College, Dublin), a book that has inspired art, regional tourism, and has become iconic of a culture, show how manuscripts can be sources of economic activity and creativity. Another example is the Rothschild Codex, a prayerbook decorated in the Flemish style. One of its iconic illustrations has been used to create an i-pad cover. Digitisation and other computer-assisted research opens up the possibility of tapping into this huge, unused cultural and economic resource to benefit society. An example is the Vernon manuscript (Bodleian Library, Oxford). A recent digitisation and research project on this manuscript has enabled it to become known to a much wider audience and to connect people with their regional literary and linguistic heritage.

Finding solutions to the problems involved in making manuscript culture more accessible is expected to have technological benefits beyond the heritage domain. Research into these problems, such as how to search digital images using computer vision methods, is tackling problems at the edge of what technology today can achieve. This work can be expected to

yield results with applications in all of the other fields where computer vision could make a difference.

5 Open Problems

The workshop participants have identified a multitude of research questions and open problems. These are itemized below and are further explored in the manifesto which resulted from the workshop.

- There are different techniques (text recognition, word-spotting, image analysis) and different questions (writer identification, classification): the question is how to make better use of them.
- Wordspotting is very appealing to cultural heritage institutions since it may prove very useful for indexing large collections, but, research remains to be done on:
  - Pre-processing of images (background and foreground)
  - Typing words, creating an ideal image of what the user is searching, and then searching in a very user friendly way (although this needs a lot of research to be carried out across collections and data-sets with very different script families).
  - Above all, taking the variability of the graphical system into account. If the end user is typing the letters, the system has to manage all allographs to find the different forms of a word.
- Prior knowledge:
  - We need to include more textual resources, so that the computer can have a better separation of words (current dictionaries are still not enough).
  - We need to combine different techniques and prior knowledge efficiently. How can we automatically align available digital images with available texts (even if not direct transcriptions)?
  - We need to create a system for aligning, overcoming textual variability, extracting forms and giving the possibility of monitoring all data at different levels (word/letter) and adding new information (abbreviations, allographs).
- Combining techniques: we need to incorporate text recognition for the alignment of images. This
  - would create a complete dictionary of all forms
  - would allow major synoptic editions
  - would create a standard data-set (much bigger than the IAM data-set)
  - would enable real research on script history
- Image analysis: creating 'mid-level' features. We need research for creating new features, inspired by human expertise
- New features:
  - Strokes (identification of different strokes; analysis of them in a palaeographically accurate way)
  - Allographs (intra-allographic and inter-allographic variability)
  - At a letter scale, to be combined with text recognition (cf. 'Efficiently combine different techniques', above, since text recognition is not a solved problem).
- Image analysis: matching existing features with visual cognition

- Doing research to interpret the features that already exist. (NB it is clear that the features do not measure what the human eye perceives, but what they perceive is probably correlated to formal phenomena that the human eye/brain can be aware of). Combining techniques and prior knowledge: use the content at the same time (identify the letters and see how they influence the measurements of features)
- Ergonomics, cognitive approach, and visualization
- Visualization of data, and presentation of data that adressess the pre-attentive perception of the researchers is not only an efficient way to promote dialog with Humanities: visualization also helps also computer science for validation during the research process. It is efficient, accurate, and offers a cross-validation since it confronts the results with another semiotic (e.g. analysis of contours, if you visualize them, you can tell better than through mathematical cross-examination if the results are possible).
- Enhanced visualization is a way to provide researchers in computer science with feedback of experts and researchers from other sciences and to support interaction of researchers
- Research remains to be done on visualization and human-computer interface and the cognitive needs to improve the comprehension of the results => user groups studies
- Hyperspectral imaging could be more efficient if a program were added into the camera to set the parameters automatically
- Visualization is also a way to efficiently introduce mid-level features

6 References

1. Tanya Clement, Sara Steger, John Unsworth, and Kirsten Uszkalo. *How Not to Read a Million Books*. http://people.lis.illinois.edu/~unsworth/hownot2read.html
2. F. Cloppet, H. Daher, V. Eglin, H. Emptoz, M. Exbrayat, G. Joutel, F. Lebourgeois, L. Martin, I. Moalla, I. Siddiqi, N. Vincent. New tools for exploring, analysing and categorising medieval scripts. *Digital Medievalist* 7, 2011. http://digitalmedievalist.org/journal/7/cloppet/
3. Tom Davis. The practice of handwriting identification. *The Library* 8:251–276, 2007. doi: 10.1093/library/8.3.251
4. Albert Derolez, *The Palaeography of Gothic Manuscript Books*, 2003. Cambridge University Press.
5. Matthieu Exbrayat and Lionel Martin. *Explorer3D*. http://www.univ-orleans.fr/lifo/software/Explorer3D/
6. David Ganz. 'Editorial palaeography': One teacher's suggestions. *Gazette du livre médiévale* 16:17–20, 1990. http://www.palaeographia.org/glm/glm.htm?art=ganz
7. Martin Jessop. Digital visualisation as scholarly activity. *Literary and Linguistic Computing* 23:281–293, 2008. doi: 10.1093/llc/fqn016
8. Lionel Martin, Matthieu Exbrayat, Guillaume Cleuziou and Fréderic Moal. Interactive and progressive constraint definition for dimensionality reduction and visualization. *Advances in Knowledge Discovery and Management* Vol. 2 (AKDM-2), pp. 121–136, 2012. Springer
9. Wendy Scase. Medieval manuscript heritage: Digital research challenges and opportunities. *Safeguard of Cultural Heritage: A Challenge from the Past for the Europe of Tomorrow: COST Strategic Workshop, 11-13 July, 2011, Florence*, pp. 97–99, 2011. Florence University Press.
10. Wendy Scase, ed. *The Vernon Manuscript: A Digital Facsimile Edition of Oxford, Bodleian Library, MS Eng.poet.a.1*, Bodleian Digital Texts 3, 2012. Oxford.

11 B. Sculley and D.M. Pasanek. Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing* 23:409-424, 2008. doi: 10.1093/llc/fqn019

12 Smith, M.H. *Les formes de l'alphabet latin, entre lécriture et lecture*, 2011. Paris.

13 Peter A. Stokes. Palaeography and image processing: Some solutions and problems. *Digital Medievalist* 3, 2007/8. http://www.digitalmedievalist.org/journal/3/stokes/

14 Peter A. Stokes. Computer-aided palaeography, present and future. In *Kodikologie und Paläographie im Digitalen Zeitalter — Codicology and Palaeography in the Digital Age*, ed. by M. Rehbein *et al.*, 2009, pp. 313-42. Books on Demand. urn:nbn:de:hbz:38-29782

15 Stutzmann, D. 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*, ed. by F. Fischer *et al.*, pp. 247–277, 2009. Books on Demand.

16 *DigiPal: Database and Resource of Palaeography, Manuscripts and Diplomatic*. http://www.digipal.eu

17 *Late Medieval English Scribes*. http://www.medievalscribes.com

18 *ManCASS C11 Database of Script and Spelling*. http://www.arts.manchester.ac.uk/mancass/C11database/

19 *Manuscripts Online: Written Culture 1000 to 1500* http://manuscriptsonline.wordpress.com/

20 *Mapping the Republic of Letters*. https://republicofletters.stanford.edu

21 *The Vernon Manuscript Project*, University of Birmingham. http://www.birmingham.ac.uk/vernonmanuscript

Participants

- Dimitris Arabadjis
National TU – Athens, GR
- Nachum Dershowitz
Tel Aviv University, IL
- Matthieu Exbrayat
Université d'Orleans, FR
- Shira Faigenbaum
Tel Aviv University, IL
- Melanie Gau
TU Wien, AT
- Tal Hassner
Open University – Israel, IL
- R. Manmatha
University of Massachusets – Amherst, US
- Ophir Münz-Manor
The Open University of Israel – Raanan, IL
- Eyal Ofek
Microsoft Res. – Redmond, US
- Micalis Panagopoulos
Ionian University – Corfu, GR
- Robert Sablatnig
TU Wien, AT
- Wendy Scase
University of Birmingham, GB
- Timothy Stinson
North Carolina State Univ., US
- Peter A. Stokes
King's College London, GB
- Dominique Stutzmann
CNRS – Paris, FR
- Ségolène Tarte
University of Oxford, GB
- Lior Wolf
Tel Aviv University, IL

Report from Dagstuhl Seminar 12391

Algorithms and Complexity for Continuous Problems

Edited by

Alexander Keller[1], Frances Kuo[2], Andreas Neuenkirch[3], and Joseph F. Traub[4]

1 **NVIDIA GmbH – Berlin, DE,** keller.alexander@gmail.com
2 **University of New South Wales, AU,** f.kuo@unsw.edu.au
3 **Universität Mannheim, DE,** neuenkirch@kiwi.math.uni-mannheim.de
4 **Columbia University, US,** traub@cs.columbia.edu

Abstract

From 23.09.12 to 28.09.12, the Dagstuhl Seminar 12391 Algorithms and Complexity for Continuous Problems was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar can be found in this report. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Seminar 23.–28. September, 2012 – www.dagstuhl.de/12391
1998 ACM Subject Classification E.1 Data Structures, F.2 Analysis of Algorithms and Problem Complexity
Keywords and phrases Computational complexity of continuous problems, Partial information, Biomedical learning problems, Random media, Computational finance, Noisy data, Tractability, Quantum computation, Computational stochastic processes, High-dimensional problems
Digital Object Identifier 10.4230/DagRep.2.9.200
Edited in cooperation with Martin Altmayer

1 Executive Summary

Alexander Keller
Frances Kuo
Andreas Neuenkirch
Joseph F. Traub

This was already the 11th Dagstuhl Seminar on Algorithms and Complexity for Continuous Problems over a period of 21 years. It brought together researchers from different communities working on computational aspects of continuous problems, including computer scientists, numerical analysts, applied and pure mathematicians. Although the seminar title has remained the same many of the topics and participants change with each seminar and each seminar in this series is of a very interdisciplinary nature.
Continuous computational problems arise in diverse areas of science and engineering. Examples include path and multivariate integration, approximation, optimization, as well as operator equations. Typically, only partial and/or noisy information is available, and the aim is to solve the problem within a given error tolerance using the minimal amount of computational resources. For example, in high-dimensional integration one wants to compute

an ε-approximation to the integral with the minimal number of function evaluations. Here it is crucial to identify first the relevant variables of the function. Understanding the complexity of such problems and construction of efficient algorithms is both important and challenging. The current seminar attracted 51 participants from more than 10 different countries all over the world. About 30% of them were young researchers including PhD students. There were 40 presentations covering in particular the following topics:

- Biomedical learning problems
- Random media
- Computational finance
- Noisy data
- Tractability
- Quantum computation
- Computational stochastic processes
- High-dimensional problems

The work of the attendants was supported by a variety of funding agencies. This includes the Deutsche Forschungsgemeinschaft, the Austrian Science Fund, the National Science Foundation (USA), and the Australian Research Council. Many of the attendants from Germany were supported within the DFG priority program SPP 1324 on "Extraction of Quantifiable Information from Complex Systems", which is strongly connected to the topics of the seminar.

As always, the excellent working conditions and friendly atmosphere provided by the Dagstuhl team have led to a rich exchange of ideas as well as a number of new collaborations. Selected papers related to this seminar will be published in a special issue of the Journal of Complexity.

2 Table of Contents

3 Overview of Talks

3.1 Quadrature of discontinuous functionals of the Heston Price using Malliavin calculus

Martin Altmayer (Universität Mannheim, DE)

Joint work of Altmayer, Martin; Neuenkirch, Andreas

The Heston Model is a popular stochastic volatility model in mathematical finance. While there exist several numerical methods to compute functionals of the Heston price, the convergence order is typically low for discontinuous functionals. In this talk, we will study an approach based on the integration by parts formula from Malliavin calculus to overcome this problem: The original function is replaced by a function involving its antiderivative and by a Malliavin weight. Using the drift-implicit Euler scheme for the square root of the volatility, we will construct an estimator for which we can prove that it has L^2-convergence order $1/2$ even for discontinuous functionals. This leads to an efficient multilevel algorithm.

3.2 Global Optimization Using Adaptive Delaunay Meshes

James M. Calvin (NJIT – Newark, US)

We describe a global optimization algorithm for twice-continuously differentiable functions that uses only adaptively chosen function evaluations. New evaluations are selected based on the function values at the vertices of the simplexes of the Delaunay triangulation of the previous evaluation points. In the case where a quality bound is available for the Delaunay triangulation, an asymptotic error bound is given.

3.3 The continuous shearlet transform in arbitrary space dimensions: general setting, shearlet coorbit spaces, traces and embeddings

Stephan Dahlke (Philipps-Universität Marburg, DE)

Joint work of Dahlke, Stephan; Steidl, Gabriele; Teschke, Gerd

In the context of directional signal analysis several approaches have been suggested such as ridgelets, curvelets, contourlets, shearlets and many others. Among all these approaches, the shearlet transform is outstanding because it is related to group theory, i.e., this transform can be derived from a square-integrable representation the so-called shearlet group. Therefore, in the context of the shearlet transform all the powerful tools of group representation theory can be exploited. Moreover, there is a very natural link to another useful concept, namely the coorbit space theory introduced by Feichtinger and Groechenig in a series of papers. By means of the coorbit space theory, it is possible to derive in a very natural way scales of smoothness spaces associated with the group representation. In this setting, the smoothness

of functions is measured by the decay of the associated shearlet transform. Moreover, by a tricky discretization of the representation, it is possible to obtain Banach frames for these smoothness spaces.

Once these new shearlet smoothness spaces are established some natural questions arise.

- How do these spaces really look like?
- What are the relations to classical smoothness spaces such as Besov spaces?
- What are the associated trace spaces?

We show that for natural subclasses of shearlet coorbit spaces which correspond to 'shearlets on the cone', there exist embeddings into homogeneous Besov spaces and that for the same subclasses, the traces onto the coordinate axis can be identified with shearlet coorbit spaces.

3.4 Complexity of Banach space valued and parametric integration

Thomas Daun (TU Kaiserslautern, DE)

Joint work of Daun, Thomas; Stefan Heinrich

We study the complexity of Banach space valued integration. The input data are assumed to be r-smooth. We consider both definite and indefinite integration and analyse the deterministic and the randomized setting. We develop algorithms, estimate their error, and prove lower bounds. In the randomized setting the optimal convergence rate turns out to be related to the geometry of the underlying Banach space.

Then we study the corresponding problems for parameter dependent scalar integration. For this purpose we use the Banach space results and develop a multilevel scheme which connects Banach space and parametric case.

3.5 Quantization by empirical measures

Steffen Dereich (Universität Münster, DE)

Main reference S. Dereich, M. Scheutzow, R. Schottstedt, "Constructive quantization: approximation by empirical measures," Ann. Inst. Henri Poincaré (B)
URL http://arxiv.org/abs/1108.5346

We study the approximation of a probability measure μ on $\mathbb{R}^d$ by its empirical measure $\hat{\mu}_N$ interpreted as a random quantization. As error criterion we consider an averaged p-th moment Wasserstein metric. In the case where $2p < d$, we establish refined upper and lower bounds for the error, a high-resolution formula. Moreover, we provide a universal estimate based on moments, a so-called Pierce type estimate. In particular, we show that quantization by empirical measures is of optimal order under weak assumptions.

3.6 On the convergence analysis of Rothe's method with spatial adaptivity

Nicolas Döhring (TU Kaiserslautern, DE)

Joint work of Cioica, Petru A.; Dahlke, Stephan; Döhring, Nicolas; Friedrich, Ulrich; Kinzel, Stefan; Lindner, Felix; Raasch, Thorsten; Ritter, Klaus; Schilling, René L.
URL http://www.dfg-spp1324.de/download/preprints/preprint124.pdf

We study the approximation of the solution u of the heat equation

$$\begin{aligned} u'(t) &= \Delta u(t) + f(u(t)) \quad \text{on } \mathcal{O},\ t \in (0,T], \\ u(0) &= u_0, \\ u|_{\partial\mathcal{O}} &= 0, \end{aligned}$$

on a bounded Lipschitz domain $\mathcal{O} \subset \mathbb{R}^d$.

First, we discretize uniformly in time using a linear implicit Euler scheme. Then, for the spatial approximation, we use an adaptive wavelet algorithm. This approach of discretizing first in time and then in space is called Rothe's method or horizontal method of lines. We illustrate convergence rates and degrees of freedom needed to obtain this rate and show how to tune the spatial approximation errors in each step.

We show that this theory can be generalized to an abstract setting, covering deterministic evolution equations

$$u'(t) = F(t,u(t)),\ t \in (0,T],\ u(0) = u_0,$$

on a separable Hilbert space V with more general S-stage schemes and semi-linear stochastic PDEs

$$du(t) = (Au(t) + f(u(t)))dt + B(u(t))dW(t),\ u(0) = u_0,$$

on a suitable function space U.

This research is supported by the DFG priority program 1324.

3.7 Sparse Stabilization and Optimal Control of the Cucker-Smole model

Massimo Fornasier (TU München, DE)

In this talk we present the stabilization and optimal control of the Cucker and Smale nonlinear dynamical system modelling consensus emergence of interacting agents. The concept of optimality of such sparse stabilization and control procedure will be discussed. This will allow us to open the question of whether the proposed strategy is really optimal in terms of its complexity.

3.8 First exit times of continuous Itô processes

Stefan Geiss (Universität Innsbruck, AT)

Joint work of Bouchard, B.; Geiss, S.; Gobet, E.

We consider in a general setting exit times from certain domains for continuous Itô-processes and their approximation by discretized processes. Applications within the simulation of BSDEs on certain domains were the starting point of the investigations.

3.9 Infinite-dimensional integration: Optimal randomized multilevel algorithms in the ANOVA setting and related new results

Michael Gnewuch (Universität Kiel, DE)

Joint work of Gnewuch, Michael; Baldeaux, Jan; Dick, Josef
Main reference J. Baldeaux, M. Gnewuch. arXiv:1209.0882v1 [math.NA]; J. Dick, M. Gnewuch. arXiv:1210.4223 [math.NA]; M. Gnewuch. arXiv:1209.1808 [math.NA]
URL http://arxiv.org/archive/math

We present new upper and lower error bounds for the infinite-dimensional numerical integration problem on weighted Hilbert spaces with norms induced by an underlying function decomposition of ANOVA or anchored type. The weights model the relative importance of different groups of variables. We have results for randomized and deterministic algorithms, and our error bounds are in both settings sharp. The upper error bounds are based on randomized or deterministic multilevel algorithms [10, 7] and/or on deterministic changing dimension algorithms [13, 15].

Let us describe our findings in more detail:

In the paper [6] we provide lower error bounds for general deterministic linear algorithms and matching upper error bounds with the help of suitable multilevel algorithms and changing dimension algorithms.

More precisely, the spaces of integrands are weighted reproducing kernel Hilbert spaces with norms induced by an anchored function space decomposition. The error criterion is the deterministic worst case error. We study two cost models [5, 13] for function evaluation which depend on the number of active variables of the chosen integration points, and two classes of weights, namely product and order-dependent (POD) weights [12] and the newly defined weights with finite active dimension. We show for both classes of weights that multilevel algorithms achieve the optimal convergence rate in one cost model while changing dimension algorithms achieve the optimal rate in the other model. In particular, we improve on results presented in [14, 8].

As an example, we discuss the infinite-dimensional anchored Sobolev space with smoothness parameter α and provide new optimal quasi-Monte Carlo multilevel algorithms and quasi-Monte Carlo changing dimension algorithms based on higher-order polynomial lattice rules [2, 3].

In [9] we consider the same spaces of integrands, but instead of deterministic algorithms and the deterministic worst-case error, we study randomized algorithms and the randomized (worst-case) error. Again, we investigate the two cost models mentioned above. We prove the first non-trivial lower error bounds for randomized algorithms in these cost models and

demonstrate their quality in the case of product weights [16]. In particular, we show that the randomized changing dimension algorithms provided in [15] achieve optimal convergence rates.

In the paper [4] we discuss the randomized ANOVA setting, which is technically more demanding. Here we focus on the cost model proposed in [5]. Our analysis refines and extends the analysis provided in [11] substantially and leads to matching upper and lower (i.e., optimal) error bounds.

As an illustrative example, we discuss the infinite-dimensional unanchored Sobolev space as space of integrands and employ randomized quasi-Monte Carlo (QMC) multilevel algorithms based on scrambled polynomial lattice rules [1].

References

1 J. Baldeaux J. Dick. A construction of Polynomial Lattice Rules with small gain coefficients. Num. Math. 119 (2011), 271–297.

2 J. Baldeaux, J. Dick, G. Leobacher, D. Nuyens and F. Pillichshammer. Efficient calculation of the worst-case error and (fast) construction component-by-component construction of higher order polynomial lattice rules. Numer. Algorithms 59 (2012), 403–431.

3 J. Baldeaux, J. Dick, J. Greslehner, F. Pillichshammer, Construction algorithms for higher order polynomial lattice rules. J. Complexity, 27, 281–299, 2011.

4 J. Baldeaux, M. Gnewuch. Optimal randomized multilevel algorithms for infinite-dimensional integration on function spaces with ANOVA-type decomposition. arXiv:1209.0882v1 [math.NA], Preprint 2012.

5 Creutzig, J., Dereich, S., Müller-Gronbach, T., Ritter, K.: Infinite-dimensional quadrature and approximation of distributions. Found. Comput. Math. 9 (2009), 391–429.

6 J. Dick, M. Gnewuch. Infinite-dimensional integration in weighted Hilbert spaces: anchored decompositions, optimal deterministic algorithms, and higher order convergence. arXiv:1210.4223 [math.NA], Preprint 2012.

7 M. B. Giles. Multilevel Monte Carlo path simulation. Oper. Res. 56 (2008), 607–617.

8 M. Gnewuch. Infinite-dimensional integration on weighted Hilbert spaces. Math. Comp. 81 (2012), 2175–2205.

9 M. Gnewuch. Lower error bounds for randomized multilevel and changing dimension algorithms. arXiv:1209.1808 [math.NA],Preprint 2012. To appear in: J. Dick, F. Y. Kuo, G. W. Peters, I. H. Sloan (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2012, Springer.

10 S. Heinrich. Monte Carlo complexity of global solution of integral equations. J. Complexity 14 (1998), 151–175.

11 F. J. Hickernell, T. Müller-Grobach, B. Niu, K. Ritter. Multi-level Monte Carlo algorithms for infinite-dimensional integration on $\mathbb{R}^{\mathbb{N}}$. J. Complexity 26 (2010), 229–254.

12 F. Y. Kuo, C. Schwab, I. H. Sloan. Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. Preprint (2011).

13 F. Y. Kuo, I. H. Sloan, G. Wasilkowski, H. Woźniakowski. Liberating the dimension. J. Complexity 26 (2010), 422–454.

14 B. Niu, F. J. Hickernell, T. Müller-Gronbach, K. Ritter. Deterministic multi-level algorithms for infinite-dimensional integration on $\mathbb{R}^{\mathbb{N}}$. J. Complexity 27 (2011), 331–351.

15 L. Plaskota, G. W. Wasilkowski. Tractability of infinite-dimensional integration in the worst case and randomized settings. J. Complexity 27 (2011), 505–518.

16 I. H. Sloan, H. Woźniakowski. When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complexity 14 (1998), 1–33.

3.10 Infinite-dimensional integration with respect to the countable product of standard normal distributions

Mario Hefter (TU Kaiserslautern, DE)

We consider a quadrature problem on the sequence space $\mathbb{R}^{\mathbb{N}}$, where the underlying measure $\mu = N(0,1)^{\mathbb{N}}$ is given by the countable product of standard normal distributions and the integrands $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ belong to a unit ball of a reproducing kernel Hilbert space $H_{\gamma,\sigma}$. The space is defined by

$$H_{\gamma,\sigma} = \bigotimes_{i\in\mathbb{N}} \left(W_\sigma^{1,2}, \|f\|_i^2 = f(0)^2 + \frac{1}{\gamma_i}\|f'\|_{L_\sigma^2}^2 \right)$$

for a sequence $\gamma = (\gamma_i)_{i\in\mathbb{N}}$ of positive weights and a positive variance parameter $\sigma > 0$. Here $W_\sigma^{1,2}$ denotes the Sobolev space of once differentiable functions, where the function itself and the derivative have bounded L^2 norm with respect to the centered normal distribution $N(0,\sigma^2)$ with variance σ^2.

We consider deterministic algorithms in the worst-case-setting, where the cost of evaluating a function at a point is the index of the highest nonzero component. Upper and lower bounds for the complexity are derived. These bounds are sharp if the sequence γ tends to zero sufficiently fast or slow and σ tends to infinity.

Motivated by an option pricing problem for a path dependent payoff function in the Black-Scholes model, we consider an integrand $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ and determine γ and σ for which $f \in H_{\gamma,\sigma}$ holds. For weights γ with $\gamma_i \asymp \frac{1}{i^\beta}$ it turns out that $f \in H_{\gamma,\sigma}$ holds if and only if no good algorithm is available on $H_{\gamma,\sigma}$.

3.11 Complexity of Banach space valued and parametric initial value problems

Stefan Heinrich (TU Kaiserslautern, DE)

Joint work of Daun, Thomas; Heinrich, Stefan

We study the complexity of initial value problems for Banach space valued ordinary differential equations both in the deterministic and in the randomized setting. The right-hand side is assumed to be r-smooth. In the randomized setting the obtained complexity estimates are related to the type of the underlying Banach space. The results extend previous ones for the finite dimensional case.

Then we apply these results to initial value problems for parameter dependent ordinary differential equations. We develop a multilevel Monte Carlo algorithm, investigate its convergence, prove matching lower bounds, and thus, settle the complexity of this problem.

3.12 The Curse of Dimensionality for Numerical Integration of Smooth Functions

Aicke Hinrichs (Universität Jena, DE)

Joint work of Hinrichs, Aicke; Ullrich, Mario; Novak, Erich; Wozniakowski, Henryk

We prove the curse of dimensionality for multivariate integration for a number of classes of smooth functions. In particular, for the class of r times continuously differentiable d-variate functions whose values are at most one the curse holds iff the bound on all derivatives up to order r does not go to zero faster than $d^{-1/2}$. We also consider the case of infinitely many differentiable functions and prove the curse if the bounds on the successive derivatives are appropriately large. The proof technique is based on a volume estimate of a neighborhood of the convex hull of n points which decays exponentially quickly if n is small relative to$d$.

3.13 Approximation of infinitely many times differentiable functions in weighted Korobov spaces

Peter Kritzer (University of Linz, AT)

Joint work of Dick, Josef; Kritzer, Peter; Pillichshammer, Friedrich; Wozniakowski, Henryk
Main reference J. Dick, P. Kritzer, F. Pillichshammer, H. Wozniakowski. Approximation of infinitely many times differentiable functions in Korobov spaces. Preprint, 2012.

We discuss L_2 approximation of functions from a weighted Korobov space of periodic infinitely many times differentiable functions for which the Fourier coefficients decay exponentially fast. We would like to check conditions on the weights such that the approximation error converges exponentially fast. Furthermore, we discuss the concepts of weak, polynomial, and strong polynomial tractability, how they are related to each other, and which properties of the weights are necessary and/or sufficient for these concepts to hold.

Research supported by the Austrian Science Fund (FWF), Project P23389- N18.

3.14 Tractability of approximation in Gaussian reproducing kernel Hilbert spaces

Thomas Kühn (University of Leipzig, DE)

Let H_d be the RKHS generated by the Gaussian kernel $K(x,y) = \exp\left(-\sum_{j=1}^{d} \sigma_j^2 (x_j - y_j)^2\right)$, where (σ_j) is a bounded seq. $(\sigma_j > 0)$ and $(x_j), (y_j) \in [0,1]^d$. We show that the approximation problem $I_d \colon H_d \to C([0,1]^d)$ is weakly tractable, if $\lim \sigma_j = 0$ and quasi-polynomially tractable, if $\sum_{j=1}^{\infty} \sigma_j^2 = \infty$.

3.15 Orthogonal transforms for QMC

Gunther Leobacher (University of Linz, AT)

Joint work of Leobacher, Gunther; Irrgeher, Christian

It has been found in the late 1990's that certain transformations of the integrand can help to increase the efficiency of quasi-Monte Carlo methods. Prominent examples from quantitative finance are provided be the Brownian bridge construction [6] and the Principal Component Analysis construction [1] for sample paths of Brownian motion.

It was later observed in [7] that

1. those transformations do not increase efficiency for arbitrary problems, rather they can slow things down for some problems;
2. those transforms can be understood as orthogonal transforms of the standard normal input vector.

In [3] the authors had the idea of constructing orthogonal transforms tailored to a given (finance) problem. Their idea has been built upon, among others, by [4, 5, 2, 8].

Up to now, most work concentrates on making the problem "as one-dimensional as possible" by choosing some orthogonal transform that puts as much variance as possible onto the first input variable.

We generalize this idea and we want to know under which conditions an orthogonal transform can be found that makes QMC more efficient, and how such a transform can be constructed.

References

1 P. Acworth, M. Broadie, and P. Glasserman. A comparison of some Monte Carlo and quasi-Monte Carlo techniques for option pricing. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1996, Proceedings of a Conference at the University of Salzburg, Austria, July 9–12, 1996*, pages 1–18, New York, 1998. Springer.

2 C. Irrgeher and G. Leobacher. Fast orthogonal transforms for pricing derivatives with quasi-Monte Carlo. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, editors, *Proceedings of the 2012 Winter Simulation Conference*, 2012. To appear.

3 J. Imai and K. S. Tan. A general dimension reduction technique for derivative pricing. *J. Comput. Finance*, 10:129–155, 2007.

4 J. Imai and K. S. Tan. An accelerating quasi-Monte Carlo method for option pricing under the generalized hyperbolic Lévy process. *SIAM J. Sci. Comput.*, 31(3):2282–2302, 2009.

5 G. Leobacher. Fast orthogonal transforms and generation of Brownian paths. *J. Complexity*, 28:278–302, 2012.

6 B. Moskowitz and R. E. Caflisch. Smoothness and dimension reduction in Quasi-Monte Carlo methods. *Math. Comput. Model.*, 23(8-9):37 – 54, 1996.

7 A. Papageorgiou. The Brownian bridge does not offer a consistent advantage in quasi-Monte Carlo integration. *J. Complexity*, 18(1):171–186, 2002.

8 I. H. Sloan and X. Wang. Quasi-Monte Carlo methods in financial engineering: An equivalence principle and dimension reduction. *Operations Research*, 59(1):80–95, 2011.

3.16 The real number PCP theorem

Klaus Meer (BTU Cottbus, DE)

Joint work of Baartse, Martijn; Meer, Klaus

In this talk we deal with probabilistically checkable proofs in the realm of real and complex number computations as introduced by Blum, Shub, and Smale. For the corresponding complexity classes $NP_{\mathbb{R}}$ and $NP_{\mathbb{C}}$, respectively, we give a characterization via PCP classes along the classical PCP theorem for the Turing machine model.

3.17 Computing Quadrature Formulas for Marginal Distributions of SDEs I

Thomas Mueller-Gronbach (Universität Passau, DE)

Joint work of Mueller-Gronbach, Thomas; Ritter, Klaus; Yaroslavtseva, Larisa

We consider the problem of approximating the marginal distribution of the solution of a stochastic differential equation (SDE) by probability measures with finite support, i.e., by quadrature formulas with positive weights summing up to one. We study deterministic algorithms in a worst case analysis with respect to classes of SDEs, which are defined in terms of smoothness constraints for the coefficients of the equation. The worst case error of an algorithm is defined in terms of a metric on the space of probability measures on the state space of the solution. We present and discuss sharp asymptotic bounds on the respective N-th minimal errors.

3.18 Learning in variable RKHSs with Application to the Blood Glucose Reading

Valeriya Naumova (RICAM – Linz, AT)

Joint work of Naumova, Valeriya; Sergei, Pereverzyev; Sivananthan, Sampath
Main reference Naumova, Valeriya; Sergei, Pereverzyev; Sivananthan, Sampath, Extrapolation in variable RKHSs with application to the blood glucose reading, Inverse Problems 27 (2011), 075010.
URL http://dx.doi.org/10.1088/0266-5611/27/7/075010

Recent progress in diabetes technology is related to the so-called continuous glucose monitoring (CGM) systems estimating the blood glucose (BG) from the electric current measured in the interstitial fluid.

In accordance with the manufacture instruction a CGM should be re-calibrated several times per day, which means that each time a drop of the blood needs to be taken to measure the actual glucose and to correct the system. Even with this procedure the system is not always accurate. Therefore, the aim is to increase the CGM accuracy.

Mathematically the problem can be formulated as follows: we are given a data set, where each element is a value of unknown function paired with a point at which this value is

attained. The function values may be blurred by noise, and the problem is to approximate the unknown function for the whole range of relevant values of the argument. This problem can be seen as an extrapolation, since it is not guaranteed that given data points span the required range.

In the context of diabetes technology, a given function value is the glucose concentration in a blood sample, and it is paired with the value of subcutaneous electric current measured at the moment when the sample is taken.

It is well-known that extrapolation is ill-posed and requires special regularization that trades off between data fitting and a complexity of a data fitter. The latter one is often measured by the norm in some reproducing kernel Hilbert space (RKHS), such as a Sobolev space, for example.

Classical regularization theory restricts itself to the case when a RKHS is assumed to be a priori known. At the same time, for a wide variety of applications a choice of RKHS is not given a priori, but should be driven by data. There are very few papers discussing this issue. Furthermore, to the best of our knowledge, no study has been reported in which the choice of RKHS is oriented towards extrapolation.

In this talk we present a new scheme of a kernel adaptive regularization learning algorithm, where the kernel and the regularization parameter are adaptively chosen within the regularization procedure. Experiments with clinical data show that the proposed choice allows essential reduction of erroneous BG-estimations as compared to commercially available CGM-devices. Moreover, the proposed approach can be used for BG-prediction and is a part of the Patent Application EP 11163219.6 filed recently by Austrian Academy of Sciences and Novo Nordisk A/S (Denmark).

References

1 Naumova, V.; Pereverzyev, S.V.; Sivananthan, S. *Extrapolation in variable RKHSs with application to the blood glucose reading.* Inverse Problems 27 (2011), 075010.

3.19 Weighted Hilbert spaces in the porous flow problem and associated numerical challenges

James Nichols (UNSW – Sydney, AU)

Joint work of Nichols, James; Kuo, Frances; Sloan, Ian; Schwab, Christoph; Scheichl, Robert; Graham, Ivan

We present recent results in the theory of applying QMC rules to integrating PDEs with random coefficients, as for example arises in the Darcy flow problem in porous media. Proving good convergence of rank-1 lattice rules requires new results for generalised spaces. We present those results, and discuss decisions motivated by the challenging numerics of this problem.

3.20 Infinite-dimensional quadratures

Dirk Nuyens (KU Leuven, BE)

Driven by the application of calculating box integrals (expected distances between points) over a Cantor set we investigate variants of the multilevel Monte Carlo algorithm on a journey for infinite-dimensional quadratures. We propose the multidigit multilevel Monte Carlo algorithm.

3.21 Approximation of linear functionals in reproducing kernel Hilbert spaces

Jens Oettershagen (Universität Bonn, DE)

Main reference F. M. Larkin, Optimal Approximation in Hilbert Spaces with Reproducing Kernel Functions, Math. Comp. 24(112), 1970

We approximate linear functionals in reproducing kernel Hilbert spaces by linear combinations of point-evaluation functionals. We generalize a result of Larkin to the multivariate setting and give necessary conditions for a set of n d-dimensional points being optimal in the sense that the worst-case error is minimal among all linear algorithms, which rely on at most function evaluations.

3.22 On the role of tractability conditions for multivariate problems

Anargyros Papageorgiou (Columbia University, US)

Joint work of Papageorgiou, Anargyros; Petras, Iasonas

We introduce a new tractability condition and the corresponding notion of κ-weak tractability. We study linear tensor product problems and show necessary and sufficient conditions for κ-weak tractability.

3.23 Repeated Phase Estimation: Approximating the ground state energy of the Schrödinger equation

Iasonas Petras (Columbia University, US)

Joint work of Papageorgiou, Anargyros; Petras, Iasonas

We demonstrate a quantum algorithm which estimates the ground state energy for convex potentials uniformly bounded by $C > 1$ with relative error $\mathcal{O}(\varepsilon)$, using

$$\mathcal{O}\left(\varepsilon^{-(3+\frac{1}{2k})} C^{4+\frac{S+k}{2k}} d^{4+\frac{3+\eta}{2k}}\right)$$

bit queries and

$$\Theta\left(\log \varepsilon^{-3}\right) + \Theta\left(\log(C^2 d^2)\right) + \Theta\left(d \log^2 \varepsilon^{-1}\right)$$

qubits, where $2kH$ is the order of the Suzuki splitting method used to simulate the exponentials and η any positive constant.

3.24 Multivariate Integration of Infinitely Many Times Differentiable Functions in Weighted Korobov Spaces

Friedrich Pillichshammer (University of Linz, AT)

Joint work of Kritzer, Peter; Pillichshammer, Friedrich; Wozniakwski, Henryk
Main reference P. Kritzer, F. Pillichshammer, H. Wozniakowski: Multivariate integration of infinitely many times differentiable functions in weighted Korobov spaces. Math. Comp, to appear
URL http://www.finanz.jku.at/index.php?id=104

We study multivariate integration in the worst-case setting for a weighted Korobov space of periodic infinitely many times differentiable functions for which the Fourier coefficients decay exponentially fast and present conditions on the weights such that we have exponential convergence with weak, polynomial and strong polynomial tractability.

3.25 Optimal approximation of SDEs with time-irregular coefficients via randomized Euler algorithm

Pawel Przybylowicz (AGH Univ. of Science & Technology-Krakow, PL)

Joint work of Przybylowicz, Pawel; Morkisz, Pawel

We investigate pointwise approximation of the solution of a scalar stochastic differential equation in the case when a drift coefficient is a Caratheodory mapping and a diffusion coefficient is only piecewise Holder continuous. It is known that under imposed assumptions and in the worst case setting the classical Euler algorithm does not converge to the solution of the equation. We give a construction of the randomized Euler scheme and investigate its error and optimality in the worst case and asymptotic setting.

Part of this talk is based on joint work with Pawel Morkisz.

References

1 A. Jentzen, A. Neuenkirch, *A random Euler scheme for Caratheodory differential equations*, J. Comp. and Appl. Math. 224 (2009), 346-359.

2 P. Przybylowicz, *Adaptive Ito-Taylor algorithm can optimally approximate the Ito integrals of singular functions*, J. Comp. and Appl. Math. 235 (2010), 203-217.

3 P. Przybylowicz, P. Morkisz, *Strong approximation of solutions of stochastic differential equations with time-irregular coefficients via randomized Euler algorithm*, submitted.

3.26 Weighted Hilbert Spaces and Integration of Functions of Infinitely Many Variables

Klaus Ritter (TU Kaiserslautern, DE)

Joint work of Gnewuch, Michael (UNSW, Sydney; supported by the German Research Foundation DFG (GN 91/3-1) and the Australian Research Council ARC); Mayer, Sebastian (Universität Bonn); Ritter, Klaus (partially supported by the DFG within Priority Program 1324)

We study two issues that arise for integration problems for functions of infinite many variables, which were first studied in [5] and which have recently been studied intensively, see, e.g., [1, 2, 3, 4, 8, 9, 10, 11] as well as [12, 13, 14] and [7, 6] for closely related problems. The setting is based on

- a reproducing kernel for functions on a domain D,
- a family of non-negative weights γ_u, where u varies over all finite subsets of $\mathbb{N}$,
- a probability measure ρ on D.

For the construction of the function space we consider the tensor product kernels

$$k_u(\mathbf{x},\mathbf{y}) = \prod_{j\in u} k(x_j, y_j)$$

with $\mathbf{x},\mathbf{y} \in D^u$, as well as the weighted superposition

$$K = \sum_u \gamma_u k_u.$$

We show that, under mild assumptions, K is a reproducing kernel on a properly chosen domain $\mathfrak{X} \subseteq D^{\mathbb{N}}$, and the quasi-reproducing kernel Hilbert space associated to K is isomorphic to the reproducing kernel Hilbert space with kernel K in a natural way. Furthermore, $H(K)$ is the orthogonal sum of the spaces $H(\gamma_u k_u)$.

Thereafter, we relate two approaches to define an integral for functions on $H(K)$, namely via a canonical representer or with respect to the product measure $\rho^{\mathbb{N}}$ on $D^{\mathbb{N}}$. In particular, we provide sufficient conditions for the two approaches to lead to the same notion of integral.

References

1 J. Baldeaux, Scrambled polynomial lattice rules for infinite-dimensional integration, in Monte Carlo and Quasi-Monte Carlo Methods 2012, L. Plaskota and H. Woźniakowski, eds., Springer, Heidelberg, 2012, pp. 255–263.

2 M. Gnewuch, Infinite-dimensional integration on weighted Hilbert spaces, Math. Comp., 81 (2012), pp. 2175–2205.

3 M. Gnewuch, Weighted geometric discrepancies and numerical integration on reproducing kernel Hilbert spaces, J. Complexity, 28 (2012), pp. 2–17.

4 F. J. Hickernell, T. Müller-Gronbach, B. Niu, and K. Ritter, Multi-level Monte Carlo algorithms for infinite dimensional integration on RN , J. Complexity, 26 (2010), pp. 229–254.
5 F. J. Hickernell and X. Wang, The error bounds and tractability of quasi-Monte Carlo algorithms in infinite dimensions., Math. Comp., 71 (2001), pp. 1641–1661.
6 F. Y. Kuo, C. Schwab, and I. H. Sloan, Multi-level quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. Preprint 2012.
7 F. Y. Kuo, C. Schwab, and I. H. Sloan, Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. Preprint 2011, to appear in SIAM J. Numer. Anal..
8 F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski, Liberating the dimension, J. Complexity, 26 (2010), pp. 422–454.
9 B. Niu and F. J. Hickernell, Monte Carlo simulation of stochastic integrals when the cost of function evaluations is dimension dependent, in Monte Carlo and Quasi-Monte Carlo Methods 2008, P. L'Ecuyer and A. B. Owen, eds., Springer, Heidelberg, 2008, pp. 545–560.
10 B. Niu, F. J. Hickernell, T. Müller-Gronbach, and K. Ritter, Deterministic multi-level algorithms for infinite-dimensional integration on RN , J. Complexity, 27 (2011), pp. 331–351.
11 L. Plaskota and G. W. Wasilkowski, Tractability of infinite-dimensional integration in the worst case and randomized settings, J. Complexity, 27 (2011), pp. 505–518.
12 G. W. Wasilkowski, Liberating the dimension for L2-approximation, J. Complexity, 28 (2012), pp. 304–319.
13 G. W. Wasilkowski and H. Woźniakowski, Liberating the dimension for function approximation, J. Complexity, 27 (2011), pp. 86–110.
14 G. W. Wasilkowski and H. Woźniakowski, Liberating the dimension for function approximation: Standard information, J. Complexity, 27 (2011), pp. 417–440.

3.27 On the slice sampler

Daniel Rudolf (Universität Jena, DE)

Joint work of Rudolf, Daniel; Łatuszyński, Krzysztof

We consider a slice sampling procedure to sample a possibly non-normalized density function. For the slice sampler, say the simple slice sampler, it is often assumed that one can sample the uniform distribution on the slices of the density. In contrast we consider slice sampler where one has a Markov chain with the correct limit distribution on every slice. The goal is to show a lower bound of the spectral gap in terms of the spectral gap of the simple slice sampler and properties of the Markov chain on the slice.

3.28 Regularization of Ill-posed Linear Equations by the Non-stationary Augmented Lagrangian Method

Otmar Scherzer (Universität Wien, AT)

Joint work of Frick, Klaus; Scherzer, Otmar
Main reference Regularization of Ill-posed Linear Equations by the Non-stationary Augmented Lagrangian Method
URL http://dx.doi.org/10.1216/JIE-2010-22-2-217

In this work we make a convergence rates analysis of the non-stationary Augmented Lagrangian Method for the solution of linear inverse problems. The motivation for the analysis is the fact that the Tikhonov-Morozov method is a special instance of the Augmented Lagrangian Method. In turn, the latter is also equivalent to iterative Bregman distance regularization, which received much attention in the imaging literature recently.

We base the analysis of the Augmented Lagrangian Method on convex duality arguments. Thereby, we can reprove some of the convergence (rates) results for the Tikhonov-Morozov Method. In addition, by the novel analysis we can prove properties of the dual variables of the Augmented Lagrangian methods. Reinterpretation of the dual variables for the Tikhonov-Morozov method gives some new convergence rates results for the linear functionals of the regularized solutions. As a benchmark for achievable convergence rates of the Augmented Lagrangian Method in the general convex context we use the results on evaluation of unbounded operators of Groetsch, which is a special instance of the Tikhonov-Morozov method.

3.29 Novel tensor formats and non-linear Galerkin approximation

Reinhold Schneider (TU Berlin)

Hierarchical Tucker tensor format (Hackbusch) and Tensor Trains (TT) (Tyrtyshnikov) have been introduced recently offering stable and robust approximation by a low order cost. In case $V = \bigotimes_{i=1}^{d} V_i$ which is proportional to d and polynomial in the ranks. We demonstrate the behavior of these ranks, depending on bilinear approximation rates and corresponding trace class norms. For many problems, which could not be handled so far, this approach can circumvent from the curse of dimensionality. We became aware, in case $V = \bigotimes_{i=1}^{d} \mathbb{C}^2$, that these formats are equivalent to tree tensor networks states and matrix product states (MPS) introduced for the treatment of quantum spin systems. Under the assumption of moderate ranks, i.e. low entanglement, this approximation enables quantum computing without quantum computers. For numerical computations, we consider the solution of quadratic optimization problems constraint by the restriction to tensors of prescribed ranks r. For approximation by elements from this highly nonlinear subset, we developed a non-linear Galerkin framework. We analyse the (open) manifold of such tensors and its projection onto the tangent space. We further derive differential equations for the gradient flow and stationary equations based on Dirac-Frenkel variational principle.

3.30 QMC Quadratures for infinite-dimensional parametric PDE problems

Christoph Schwab (ETH Zürich, CH)

Joint work of Schwab, Christoph; Kuo, Frances; Sloan, Ian
Main reference Frances Kuo and Christoph Schwab and Ian Sloan: Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients ,SIAM. J. Numer. Anal. 2013 (accepted for publication)
URL http://www.sam.math.ethz.ch/reports/2011/52

We present recent results on well-posedness and regularity for several classes of infinite-dimensional, parametric ordinary and partial differential equation (PDE) problems. Both, linear and nonlinear PDE problems are covered.

Such problems arise, among others, in connection with diffusion, vibration and wave-propagation in random media, when a parametric representation of the law of the random solutions is sought in in terms of countably many parameters of the problems' random inputs, e.g. in terms of a Karhunen-Loeve expansion.

We present recent results on the regularity of these parametric representations of random fields, in terms of weighted reproducing kernel Hilbert spaces in infinite dimension.

We then show how recent results on Quasi Monte-Carlo quadratures on these weighted reproducing kernel Hilbert spaces in infinite dimension allow for the efficient numerical evaluation of mean fields and statistical moments of the random solutions.

The QMC results are joint work with Frances Kuo and Ian Sloan, of UNSW, Sydney, Australia.

The results are part of the research reports
http://www.sam.math.ethz.ch/reports/2012/18
http://www.sam.math.ethz.ch/reports/2012/25
http://www.sam.math.ethz.ch/reports/2011/52

3.31 Lattice Methods with Designed Weights for PDE with Random Coefficients

Ian Sloan (University of New South Wales, AU)

Joint work of Kuo, Frances (UNSW); Graham, Ivan (Bath); Nichols, James (UNSW); Scheichl, Rob (Bath); Schwab, Christoph (ETH); Sloan, Ian (UNSW)

In this talk I will present recent developments on applying quasi-Monte Carlo methods (specifically lattice methods) to the computation of high-dimensional expected values (treated as multivariate integrals) of functionals of the solution of a PDE with random coefficients. A guiding example is the flow of a liquid through a porous material, with the permeability modelled as a random field. In this work we apply the theory of lattice methods in weighted spaces, together with estimates derived from the PDE, to design special lattice methods with proved good convergence properties for the computed expected values.

3.32 Adaptive (piecewise) tensor product wavelet Galerkin method

Rob Stevenson (University of Amsterdam, NL)

In this talk we summarize the convergence theory of adaptive wavelet Galerkin methods for solving well-posed linear or nonlinear operator equations. We discuss the application of these methods with the use of (piecewise) tensor product bases. Finally, we focus on the adaptive solution of simultaneous space-time variational formulations of evolution problems.

3.33 A rapidly mixing Markov chain for the two-dimensional Ising model

Mario Ullrich (Universität Jena, DE)

Main reference M. Ullrich, Rapid mixing of Swendsen-Wang and single-bond dynamics in two dimensions
URL http://arxiv.org/abs/1202.6321

We prove that the Swendsen-Wang dynamics (SW) for the Ising model on the two-dimensional square lattice is rapidly mixing at all temperatures. For this, we present three comparison results. First we show rapid mixing at and above the critical temperature by comparison with the single-site heat- bath dynamics. Then we prove that rapid mixing of SW and the single-bond dynamics (SB) for the corresponding random-cluster model is equivalent. And finally, we relate the mixing properties of SB at high and low temperatures (using dual graphs).

3.34 Optimal Cubature in Sobolev-Besov Spaces with Dominating Mixed Smoothness

Tino Ullrich (Universität Bonn, DE)

Main reference Tino Ullrich, Optimal Cubature in Besov Spaces With Dominating Mixed Smoothness on the Unit Square, Preprint, Bonn 2012

We present new constructive and asymptotically optimal error bounds for numerical integration in bivariate periodic Besov spaces with dominating mixed smoothness $S^r_{p,q}B(T^2)$, where $1 \leq p, q \leq \infty$ and $r > 1/p$. Our first result uses Quasi-Monte Carlo integration on Fibonacci lattice rules and improves on the so far best known upper bound achieved by using cubature formula taking function values from a Sparse Grid. It is well known that there is no proper counterpart for Fibonacci lattice rules in higher dimensions. To this end, our second result is based on Hammersley (or Van der Corput) type point grids. Instead of exploiting a Hlawka-Zaremba type discrepancy duality, which is limited to small smoothness parameters $1/p < r \leq 1$, we extend Hinrichs' recent results to larger orders r, namely $1/p < r < 2$. This direct approach is strongly conjectured to have a proper counterpart for higher orders r and, in addition, for functions on the d-torus T^d . Last, but not least, we prove that any cubature rule based an a sparse grid in d dimensions has a significantly worse error order than the

previously described methods. These results are a first step to approach the problem of optimal recovery of functions from a discrete set of function values in a completely new way.

3.35 Learning Functions of Few Arbitrary Linear Parameters in High Dimensions

Jan Vybiral (TU Berlin, DE)

Joint work of Fornasier, Massimo; Schnass, Karin; Vybiral, Jan
Main reference Found. Comput. Math. 12 (2) (2012), 229-262
URL http://dx.doi.org/10.1007/s10208-012-9115-y

We study the uniform approximation of functions of many variables with the following inner structure. We assume, that $f(x) = g(Ax)$, where $x \in \mathbb{R}^d$, A is a $k \times d$ matrix and g is a (smooth) function on $\mathbb{R}^k$. Both g and A are unknown and their recovery is a part of the problem.

Under certain smoothness and variation assumptions on the function g, and an arbitrary choice of the matrix A, we present a sampling choice of the points drawn at random for each function approximation and algorithms for computing the approximating function. Due to the arbitrariness of A, the choice of the sampling points will be according to suitable random distributions and our results hold with overwhelming probability. Our approach uses tools taken from the compressed sensing framework, recent Chernoff bounds for sums of positive-semidefinite matrices, and classical stability bounds for invariant subspaces of singular value decompositions.

3.36 Average Case Tractability of Approximating ∞-Variate Functions

Grzegorz Wasilkowski (University of Kentucky, US)

We discuss function approximation in the average case setting for spaces of ∞-variate functions that have a weighted tensor product form and are endowed with a Gaussian measure that also has a weighted tensor product form. We assume that the cost of function evaluation depends on the number of active variables and we allow it to be from linear to exponential. We provide a necessary and sufficient condition for the problem to be polynomially tractable and derive the exact value of the tractability exponent. In particular, the approximation problem is polynomially tractable under modest conditions on weights even if the function evaluation cost is exponential in the number of active variables. The problem is weakly tractable even if this cost is doubly exponential.

3.37 Probabilistic star discrepancy bounds for double infinite random matrices

Markus Weimar (Universität Jena, DE)

In 2001 Heinrich, Novak, Wasilkowski and Wozniakowski proved that the inverse of the discrepancy depends linearly on the dimension by showing that a random point set P of N points in the s-dimensional unit cube satisfies the discrepancy bound $D_N^{*s}(P) < cs^{1/2}N^{-1/2}$ with positive probability. Later their results were generalized by Dick to the case of double infinite random matrices.

In this talk we give explicit, asymptotically optimal bounds for the star discrepancy of such random matrices, and give estimates for the corresponding probabilities. Using the same techniques we derive similar discrepancy bounds for randomly generated completely uniformly distributed (c.u.d.) sequences which find applications in Markov Chain Monte Carlo.

The talk is based on a recent paper which is joint work with C. Aistleitner [1].

References

1 C. Aistleitner and M. Weimar. *Probabilistic star discrepancy bounds for double infinite random matrices*. submitted manuscript, 2012 (http://users.minet.uni-jena.de/ weimar/)

3.38 Tractabilty of multi-parametric Euler and Wiener integrated processes

Henryk Woźniakowski (Columbia University, US and University of Warsow, PL)

Joint work of Lifshitz M. (State University of Saint Petersburg); Papageorgiou, Anargyros (Columbia University); Henryk Woźniakowski

We provide necessary and sufficient conditions on weak, polynomial and strong polynomial tractability for multivariate approximation in the average case setting for Gaussian measures with Euler and Wiener integrated covariance kernels. These conditions are expressd in terms of the sequence $\{r_k\}$, where r_k measures the smoothness of functions with respect to the k-th variable.

3.39 Computing Quadrature Formulas for Marginal Distributions of SDEs II

Larisa Yaroslavtseva (Universität Passau, DE)

Joint work of Yaroslavtseva, Larisa; Mueller-Gronbach, Thomas

This talk is a continuation of the talk of Thomas Mueller-Gronbach. Here we consider the case of scalar SDEs with bounded coefficients that are 6 times continuously differentiable and have bounded derivatives. Furthermore, the diffusion coefficient is assumed to be bounded away

from zero. For the definition of the error we employ the Wasserstein distance. We present a deterministic algorithm, which is based on sparse discrete approximations of Wagner-Platen steps with support points in small grids and is easy to implement. The method achieves the optimal order of convergence in terms of the computational cost, up to an arbitrarily small power of the cost.

3.40 Pointwise approximation for additive random fields

Marguerite Zani (Université Paris-Est Créteil, FR)

Joint work of Lifshitz, Mikhail A. (St. Petersburg); Zani, Marguerite

We consider standard information in the average case setting for additive random fields. We consider a multilevel algorithm using function evaluations and show that the L^2 approximation error is somehow comparable to the one in the linear case.

Participants

- Christoph Aistleitner
TU Graz, AT
- Martin Altmayer
Universität Mannheim, DE
- James M. Calvin
NJIT – Newark, US
- Ronald Cools
KU Leuven, BE
- Stephan Dahlke
Universität Marburg, DE
- Thomas Daun
TU Kaiserslautern, DE
- Steffen Dereich
Universität Münster, DE
- Nicolas Döhring
TU Kaiserslautern, DE
- Massimo Fornasier
TU München, DE
- Stefan Geiss
Universität Innsbruck, AT
- Michael Gnewuch
Universität Kiel, DE
- Mario Hefter
TU Kaiserslautern, DE
- Stefan Heinrich
TU Kaiserslautern, DE
- Aicke Hinrichs
Universität Jena, DE
- Peter Kritzer
University of Linz, AT
- Thomas Kühn
Universität Leipzig, DE
- Frances Kuo
UNSW – Sydney, AU
- Gunther Leobacher
University of Linz, AT
- Peter Mathé
Weierstraß Institut – Berlin, DE
- Klaus Meer
BTU Cottbus, DE
- Thomas Müller-Gronbach
Universität Passau, DE
- Valeriya Naumova
RICAM – Linz, AT
- James Nichols
UNSW – Sydney, AU
- Erich Novak
Universität Jena, DE
- Dirk Nuyens
KU Leuven, BE
- Jens Oettershagen
Universität Bonn, DE
- Anargyros Papageorgiou
Columbia University, US
- Sergei V. Pereverzev
RICAM – Linz, AT
- Iasonas Petras
Columbia University, US
- Friedrich Pillichshammer
University of Linz, AT
- Leszek Plaskota
University of Warsaw, PL
- Pawel Przybylowicz
AGH Univ. of Science & Technology-Krakow, PL
- Klaus Ritter
TU Kaiserslautern, DE
- Daniel Rudolf
Universität Jena, DE
- Otmar Scherzer
Universität Wien, AT
- Wolfgang Ch. Schmid
Universität Salzburg, AT
- Reinhold Schneider
TU Berlin, DE
- Christoph Schwab
ETH Zürich, CH
- Winfried Sickel
Universität Jena, DE
- Ian H. Sloan
UNSW – Sydney, AU
- Rob Stevenson
University of Amsterdam, NL
- Shu Tezuka
Kyushu University, JP
- Joseph F. Traub
Columbia University, US
- Mario Ullrich
Universität Jena, DE
- Tino Ullrich
Universität Bonn, DE
- Jan Vybiral
TU Berlin, DE
- Grzegorz Wasilkowski
University of Kentucky, US
- Markus Weimar
Universität Jena, DE
- Henryk Wozniakowski
Columbia University, US
- Larisa Yaroslavtseva
Universität Passau, DE
- Marguerite Zani
Université Paris-Est Créteil, FR