



DAGSTUHL  
MANIFESTOS

**Volume 2, Issue 1, January – December 2012**

Towards A Multi-Discipline Network Perspective (Dagstuhl Perspectives Workshop 12182) <i>Matthias Häsel, Thorsten Quandt, and Gottfried Vossen</i> .....	1
Computation and Palaeography: Potentials and Limits (Dagstuhl Perspectives Workshop 12382) <i>Tal Hassner, Malte Rehbein, Peter A. Stokes, and Lior Wolf</i> .....	14

ISSN 2193-2433

*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/dagman>

*Publication date*

July, 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>.

*License*

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



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

- Attribution: The work must be attributed to its authors.
- No derivation: It is not allowed to alter or transform this work.

The copyright is retained by the corresponding authors.

*Aims and Scope*

The manifestos from Dagstuhl Perspectives Workshops are published in the *Dagstuhl Manifestos* journal. Each manifesto aims for describing the state-of-the-art in a field along with its shortcomings and strengths. Based on this, position statements and perspectives for the future are illustrated. A manifesto typically has a less technical character; instead it provides guidelines and roadmaps for a sustainable organisation of future progress.

*Editorial Board*

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

*Editorial Office*

Roswitha Bardohl (*Managing Editor*)  
Marc Herbstritt (*Head of Editorial Office*)  
Jutka Gasiorowski (*Editorial Assistance*)  
Thomas Schillo (*Technical Assistance*)

*Contact*

Schloss Dagstuhl – Leibniz-Zentrum für Informatik  
Dagstuhl Manifestos, Editorial Office  
Oktavie-Allee, 66687 Wadern, Germany  
[publishing@dagstuhl.de](mailto:publishing@dagstuhl.de)

# Towards A Multi-Discipline Network Perspective

Edited by

Matthias Häsel<sup>1</sup>, Thorsten Quandt<sup>2</sup>, and Gottfried Vossen<sup>3</sup>

1 Otto Group – Hamburg, DE, [Matthias.Haesel@otto.de](mailto:Matthias.Haesel@otto.de)

2 Universität Hohenheim, DE, [thorsten.quandt@uni-hohenheim.de](mailto:thorsten.quandt@uni-hohenheim.de)

3 Universität Münster, DE, [vossen@uni-muenster.de](mailto:vossen@uni-muenster.de)

---

## Abstract

This is the manifesto of Dagstuhl Perspectives Workshop 12182 on a multi-discipline perspective on networks. The information society is shaped by an increasing presence of networks in various manifestations, most notably computer networks, supply-chain networks, and social networks, but also business networks, administrative networks, or political networks. Online networks nowadays connect people all around the world at day and night, and allow to communicate and to work collaboratively and efficiently. What has been a commodity in the private as well as in the enterprise sectors independently for quite some time now is currently growing together at an increasing pace. As a consequence, the time has come for the relevant sciences, including computer science, information systems, social sciences, economics, communication sciences, and others, to give up their traditional “silo-style” thinking and enter into borderless dialogue and interaction. The purpose of this Manifesto is to review where we stand today, and to outline directions in which we urgently need to move, in terms of both research and teaching, but also in terms of funding.

**Perspectives Workshop** 02.–04. May, 2012 – [www.dagstuhl.de/12182](http://www.dagstuhl.de/12182)

**1998 ACM Subject Classification** A.0 General, A.2 Reference, H. Information Systems, J.4 Social and Behavioral Sciences, K.4 Computers and Society

**Keywords and phrases** Networks, network infrastructure, network types, network effects, data in networks, social networks, social media, crowdsourcing

**Digital Object Identifier** 10.4230/DagMan.2.1.1

## Executive Summary

The information society is shaped by an increasing presence of networks in various manifestations. Efficient computer networks are regarded as a significant enabler for the process of change towards networks of any size and complexity. They serve as an administrative and technological basis for social network structures, with the result that online networks connect people all around the world at day and night, and allow to communicate and to work collaboratively, efficiently, and without recognizable time delay. Companies reduce their in-house production depth, join forces in supply chain networks and establish cooperation with their suppliers, with their customers, and even with their competitors. By now, social networks like Facebook, Google+, LinkedIn or XING are seen as the de facto standard of “social networking” in the information society. Companies are mimicking their effects internally, allow overlays of networking applications with regular business ones, and a use of social networks for enterprise purposes including and beyond advertising has become common. Public administrations create and improve shared services and establish “Private Public Partnerships (PPP)” to benefit from synergetic effects of cooperation with private and public organizations.



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

A Multi-Discipline Network Perspective, *Dagstuhl Manifestos*, Vol. 2, Issue 1, pp. 1–13

Editors: Matthias Häsel and Thorsten Quandt and Gottfried Vossen




Dagstuhl Manifestos

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

As the interactions between people in these networks increase at various levels, new approaches are needed to analyze and study networks and their effects in such a way that individuals as well as organizations and enterprises can benefit from them. Indeed, more interaction and collaboration between fields such as information systems, computer science, social sciences, economics, communication sciences and others is needed in the future in order to understand the many networks effects as well as to be able to master them appropriately. These fields need to identify a common level of language, tools and set of methodologies so that the various aspects of networking can be addressed and jointly developed further. The most important point is the need for a renewed multi-disciplinarity. To a great extent, networks are driven and further developed by practitioners, which also means that they are evolving in a very fast manner and not emanating from a single scientific discipline.

As a major result from the workshop, the following problems and directions have been identified:

1. To be able to both understand networks and their effects as well as to contribute to the state of art, true inter- or multi-disciplinary research is needed that involves the various fields mentioned above.
2. As the aforementioned disciplines grow together and embark on collaborative research, it is important to convince funding agencies that multi-disciplinary research should arrive on their agendas.
3. Web sciences need to be developed as a field, and also need to be integrated into teaching. This will most likely lead to novel curricula which receive their content from multiple disciplines in a balanced way.

 **Table of Contents**

<b>Executive Summary</b> . . . . .	1
<b>Introduction</b> . . . . .	4
<b>Data in Networks</b> . . . . .	4
<b>Network Infrastructures</b> . . . . .	6
<b>The Specifics of Social Networks and Social Media</b> . . . . .	7
<b>The (Observable) Network Effects of Crowdsourcing</b> . . . . .	8
<b>Conclusions, Findings, Recommendations</b> . . . . .	9
<b>Participants</b> . . . . .	12
<b>References</b> . . . . .	12

## 1 Introduction

The information society is shaped by an increasing presence of networks in various manifestations. Efficient computer networks are regarded as a significant enabler for the process of change towards networks of any size and complexity. They serve as an administrative and technological basis for social network structures, with the result that online networks connect people all around the world at day and night, and allow to communicate and to work collaboratively, efficiently, and without recognizable time delay. Companies reduce their in-house production depth, join forces in supply chain networks and establish cooperation with their suppliers, with their customers, and even with their competitors. By now, social networks like Facebook, Google+, LinkedIn or XING are seen as the de facto standard of “social networking” in the information society. Companies are mimicking their effects internally, allow overlays of networking applications with regular business ones, and a use of social networks for enterprise purposes including and beyond advertising has become common. Public administrations create and improve shared services and establish “Private Public Partnerships (PPP)” to benefit from synergetic effects of cooperation with private and public organizations.

With the workshop that has led to this Manifesto, it has been our intention to focus on three fundamental aspects of networks in order to analyze and study the design, interplay, and behavior of networks in the Information Society:

1. *Drivers*: Networks can be regarded as systems that are continuously shaped by their environment. In fact, the emergent structure and properties of networks are subject to self-organizing processes — not unlike evolutionary processes — that create structure in the form of temporarily stable patterns of interaction between actors.
2. *Cohesion*: In a general context cohesion describes the phenomenon of (economic and/or social) solidarity, or, in other words, the intention of actors to act in the middle of their neighbors. Structural cohesion is the sociological and graph-theoretical conception for evaluating the behavior of social groups and networks.
3. *Dynamics*: A dynamic system is a system that changes its state over time. Concerning different network application areas, we regard the dynamics of a system as the change of states a system takes. On the one hand, we consider a change of state in a network as the exchange of entities (information, goods, etc.) between its actors. On the other, the change of state in a network is regarded as its evaluation, which may involve, among other aspects, a change of the underlying system’s structure over time.

Four distinct areas that pertain to networks and networking appear to be of particular importance and interest:

1. Data in Networks,
2. network infrastructures,
3. the specifics of social networks and social media, and
4. the (observable) network effects of crowdsourcing.

We next look at each area in turn.

## 2 Data in Networks

Networks produce massive amounts of data, either automatically through machines (e.g., Web server logs, supply-chain control) or through user input. Indeed, user-generated content has been one of the distinctive features of “Web 2.0” or the evolution of the Web from a

“read-only Web” to a “read-write Web” [13]. Moreover, accessible data on the Web, whether created by computers, by Web users, or generated within professional organizations, are growing at a tremendous pace. Social networks like Facebook, search engines like Google, or e-commerce sites like Amazon store new data in the terabyte range on a daily basis. Due to the emerging usage of cloud computing, this trend will not only continue, but accelerate over the coming years, as not only more and more data is generated, but also more and more data is permanently stored online, is linked to other data, and is aggregated in order to form new data.

Regarding the various kinds of data on the Web and in networks, including linked open data, socio-economic data, big data, and user-supplied data, relevant topics are technical aspects of data, usage patterns of data, types of data in networks (e.g., process data). Questions to be asked include, but are not limited to the following: Is storing all this data necessary? What can be done with all this data? How can data flow between networks? How can data produced in one network be beneficial for another?

Regarding data arising in the context of computer networks, a first observation was that the term “big data” should rather be “*broad data*,” as various developments, including linked data, the Web of data, and others are currently coming together. In particular linked data [3] has gained recent popularity in the context of the Semantic Web [2], as Semantic Web people think in terms of *links*, as opposed the previous thinking in terms of *pages*. This perception nowadays also applies to data creation, updating, and analysis. Surprisingly, data scraping is still in wide use, since linking is not yet fully understood and reasonable alternatives are not available (e.g., based on metadata standard formats). On the other hand, data is most useful when it can be combined with other data, which is what we currently see on social networks like Facebook with their underlying graph databases, where there is a rising usage of inherent semantics as well as implicit context.

Clearly, data is heavily spreading across networks, but we still do not understand how to create networks appropriate for a specific purpose, where the spreading of data can be directed and controlled in some way, or how to bring together (structured, semi-structured etc.) data and information. Besides data, it is important to distinguish *networks of machines* from *networks of people*: There are attempts which try to join the two, while others want to keep them apart. Human actors are obviously important in the network picture, since they are prime suppliers of data and its connections, especially in social networks.

Whether data is linked or not, what is of increasing importance is to be able to identify *data provenance* (or data lineage, i.e., the ability to trace given data to its roots, points of creation, and along its history of changes). Provenance has a different meaning for data (“where does it come from?”) and for networks; in the latter case, it is more process- or document-oriented. Data provenance [5] [6] has originated from scientific applications, e.g., in physics or in molecular biology, where reproducibility has always been an important aspect; however, provenance has meanwhile reached even business areas. From a technical perspective, provenance is often seen in connection to data curation, as exemplified, for instance, in the DBWiki project [4].

The traditional database approach to large data collections has been the *data warehouse* [10], where data is collected from various sources, put through an ETL process (short for *extraction, transformation, and loading*) and finally integrated into a single data collection, the warehouse. The latter then forms the basis for data analysis, online analytical processing (OLAP), as well as data mining. If that is to happen on the fly, a data warehouse is not good enough, since an ETL process takes too much time. A data warehouse stands for slow integration with high quality, whereas fast integration with lower quality is often more

desirable, in particular since data changes occur on the network, “by themselves.” For that, a linking of data sets seems again more appropriate; networks of data are needed, which at best amounts to a warehouse in a cloud-like world. This is in a way similar to developments in software engineering (from traditional to agile approaches).

Another important aspect is that data is increasingly considered as *goods* which have a *value* or come at a price: If the goods are rare, you collect them; if there is abundance, collecting is no longer necessary (an example is music, in particular records vs. music obtained from the Web). Indeed, marketplaces for data are on the rise [12], which aim at the development of reliable and trusted platforms for the production, provision, and use of data. In this area, where sophisticated search and analysis tools are needed, there is a link to *crowdsourcing*, i.e., the idea to outsource a task to a possibly anonymous group of people. Numerous examples from recent years prove that having the user in the loop can improve data quality (e.g., maps of Haiti before and after the quake; the UK map of bus stops before and after it had been opened to the public). However, the effects achievable with crowdsourcing depend on the specifics of the crowd. Here it is important to distinguish between a “pre-defined” crowd in a professional environment (a “club,” e.g., for building a plane) and a “randomly gathered” crowd (as in the case of the bus stops). The techniques used in either category may or may not be the same; the size of the crowd may be a determining factor: As the crowd gets larger, the need for individual experts potentially decreases, but control remains an issue and beyond a certain size experts are needed again for helping to separate useful information from nonsense. So a question is how a crowd can be triggered to do what it is expected to do or what a system (e.g., Wikipedia) requires them to do. Ultimately, such a crowd will decide about what is right and what is wrong.

### 3 Network Infrastructures

Network infrastructures increasingly shape modern societies. In comparison to traditional infrastructures such as traffic, energy or health care, network infrastructures based on the Internet and its services are developed much faster, at a considerably wider scale, and they facilitate widespread participation. Computerized network infrastructures are easily scalable due to the availability of massive computing and storage power as well as network bandwidth and due to the availability of standardized protocols. They are versatile and represent a generative regime (they facilitate the growth of new infrastructures). Several network infrastructures can thus be conceptualized as commodities and are seen as a societal resource for innovation, economic development and welfare.

Topics to be discussed in this area include decentralized network architectures, cloud computing, emergence and design of network infrastructures, simulation of network behavior, informed logistics infrastructures. Questions include the following: Which infrastructures are particularly suited for which area (e.g., SCM and logistics, service industry)? Do we still need to care about infrastructure, or will it soon be all invisible like electrical current?

The *definition* of a network infrastructure should cover aspects such as non-rivalry access, one infrastructure for one purpose, and visibility only in the case of failure. According to Nicholas Carr,<sup>1</sup> infrastructure does not really make a difference, at least as far as IT infrastructure is concerned: If each enterprise has it, it can no longer help to sustain a competitive

---

<sup>1</sup> <http://www.nicholasgarr.com/doesitmatter.html>



advantage. Important are standards; there is a technology stack with infrastructure at the lowest level. Infrastructures require administration, (legal) regulation, and accessibility. The proliferation of the network society may have an effect on infrastructures. Research topics to be studied include governance, comparison of infrastructure types, infrastructure lifecycles, vulnerability of infrastructures, as well as privacy. Twitter and Facebook have the potential to become infrastructures.

#### 4 The Specifics of Social Networks and Social Media

Social networks are at the heart of modern network usage, demonstrated by the wide user coverage (if Facebook was a country, it would currently be the third biggest in the world). They have different foci, be it on personal or professional issues (or a mixture of both), they serve as extremely efficient and sometimes highly specialized news and communication platforms (think, for example, of the role of Twitter in the Arab spring of early 2011), and they are to an increasing degree discovered by enterprises as an instrument for reaching out internally to employees and externally to customers. The result is an increasing professional investment in social media technology and advertising, although the ultimate effects, in particular the external ones, still remain to be seen.

Relevant topics in this area include (social) network analysis, social networks for the public domain, social media (networks), and social commerce. Questions to be asked are: Which distinctions can currently be made between various social networks (e.g., Facebook, Google+, Path et al.)? How could the future of social networks look like? Will Facebook be the new “operating system” of the Internet? What value do online social networks have for an economy from a macro-economic perspective? What influence does my online social network have on me, and what influence do I have as a node in that network? Can I influence my personality by forming specific (online) relationships? Is the Internet a special case for all existing research results on social networks? What are the specifics of online social networks that the social sciences provide? Does the Internet enhance existing or enable new social behavior? Does the mere size of a network or community make possible new effects that have not been possible before due to quantitative thresholds? What are parallels (and metaphors to describe them) between the real world and the online world?

It is obvious that the social sciences know a lot about social networks, but miss the technical expertise, a fact that needs to change (see the findings in Section 6). Yet the question is how social scientists (who have questions) can be brought together with information systems researchers (who have tools to answer these questions). What social sciences *can* contribute and study are questions like “what influence does my network have on me?” or “Can I form my personality through an architecture of social contacts?” Some people claim that the Internet as well as mobile devices fundamentally change the behavior of people and the way they communicate, and that the Internet is hence not just “yet another medium.”<sup>2</sup>

For online networks as they exist today and the transparency they provide (which is much larger than it used to be prior to online networks), control and regulation are needed, yet how to do this (if it can be done at all) is still vastly unclear. Governance approaches are also needed for commercial networks. On the other hand, the added value of online social networks is undeniable.

The added value of online social networks can be discussed in multiple dimensions: The

<sup>2</sup> [http://www.theshallowsbook.com/nicholascarr/Nicholas\\_Carrs\\_The\\_Shallows.html](http://www.theshallowsbook.com/nicholascarr/Nicholas_Carrs_The_Shallows.html)

personal value of users, the commercialization value generated by platform providers (such as Facebook), and the value generated by businesses (such as brands that advertise on Facebook). However, from a provider perspective, commercialization does currently focus on simply-targeted advertisements. New business models in terms of bringing together supply and demand will appear and need to be researched in the future. For example, being able to develop social software using APIs such as the Facebook Platform or OpenSocial opens up a wealth of different business opportunities because businesses do not have to build a new social graph from scratch. Internet companies can exploit this, for instance, to boost their outreach and profile immensely — by positioning their existing product on other networking sites as a social application [8].

For now, a big question for businesses is how to attribute revenues from social media to the different touchpoints a customer has had with their brand or products. People tend to switch between distinct contexts all the time (reading email on their smartphones, browsing the Web, shopping online, participating in a chat, checking in on Foursquare etc.), but the interplay of digital touchpoints still needs to get on the research agenda. Also, it is largely unclear how user behaviour in social networks depends on the device used (e.g., PC, smartphone, or tablet) and context of usage.

Physical presence has a distinct meaning in a social context. The same applies to shopping: Online shopping is different from going to a store. For example, it is usually more focused, yet there may be more impulses. *Servicescape* is a concept that was developed by Booms and Bitner to emphasize the impact of the physical environment in which a service process takes place.<sup>3</sup> As they state in their paper, “the ability of the physical environment to influence behaviors and to create an image is particularly apparent for service businesses such as hotels, restaurants, professional offices, banks, retail stores, and hospitals.” But what holds for the physical world may as well apply to the virtual world; in other words, how can I find out that the world that Google is showing me is the real world?

## 5 The (Observable) Network Effects of Crowdsourcing

One of the most striking “network effects,” besides the creation of large friends networks, is the already mentioned area of *crowdsourcing*. According to [1], “Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.” Crowdsourcing has been successfully applied to tasks that are easier to solve for a human than for a computer (e.g., image analysis), but also to many other areas, and it has meanwhile developed “subareas” such as *crowdfunding* or *crowdvoting*.

---

<sup>3</sup> <http://www.jstor.org/stable/1252042>  
<http://en.wikipedia.org/wiki/Servicescape>

Topics in this field include large-scale cooperation, collaborative editing, constructivism via digital means, knowledge management, IT supported collaboration in logistics networks, and agent-based coordination. Questions here are the following: Which effects can be observed by employing crowdsourcing? In which areas has crowdsourcing failed up to now and why? Which new areas could benefit from crowdsourcing (technical ones such as query optimization, social ones such as crowdfunding)?

Collaboration between people is often seen as a form of art (e.g., in music<sup>4</sup> or social writing) and differs from the kind of collaboration we practice today. Most examples of crowdsourcing we see today are those which are working well. Different forms of crowdsourcing have had different successes (e.g., Galaxy Zoo<sup>5</sup> and Zooniverse<sup>6</sup>), and it turns out that creative work requires a selection process for the crowd. This might lead to a revised notion of “crowd,” e.g., “active,” “participatory,” or “unconsciously voluntary:” A member may be invited, participate actively, register by herself, or be used without knowing about it.<sup>7</sup> Data querying and analysis are increasingly seen as an application for crowdsourcing (a mixture of human computation and automation) [7]. Crowdsourcing is constantly producing process data and content data. Another emerging specialization that might arise in the future is the *private crowd* (e.g., inside an enterprise) vs. the *public crowd* (e.g., AWS Mechanical Turk).

Social sciences should take a leading role in the design of social networks, not just analyze them. The goal should be to make it easier for people to meet, get together, and let them do the rest themselves. For example, flashmobs minimize risk and maximize outcome; they are a low-level form of crowdsourcing.

There are also cases where the crowd is less efficient than a hierarchy. Examples include warfare (but there is also network-centric warfare) or emergencies. On the other hand, even social media might be designed in such a way that they incorporate hierarchy (such as Wikipedia). In some areas or applications crowdsourcing is not just inapplicable, but has failed or is (or has become) inappropriate, e.g., teaching material in a university program, military applications based on classified information, or generally applications requiring fast decisions. Crowdsourcing is not even useful in an arbitrary application, since it may destroy creativity or a vision in a given setting.

## 6 Conclusions, Findings, Recommendations

More interaction and collaboration between the various fields pertinent to networks is needed. The fields need to identify a common level of language, tools and set of methodologies so that the various aspects of networking we discussed can be addressed and jointly developed further. Indeed, the most important point in our findings was the need for a renewed multi-disciplinarity. To a great extent, networks are driven and further developed by practitioners, which also means that they are evolving in a very fast manner and not emanating from a single scientific discipline. To be able to both understand them and contribute to the state of art, we need true inter- or multi-disciplinary research that involves computer science, social

<sup>4</sup> See <http://www.npr.org/2012/05/13/151712146/first-listen-hilary-hahn-and-hauschka-silfra> for an example representing the taste of one the authors and <http://www.inc.com/articles/201103/ted-collaborative-communication-social-media-age.html> for a more general coverage.

<sup>5</sup> <http://www.galaxyzoo.org/>

<sup>6</sup> <https://www.zooniverse.org/>

<sup>7</sup> Matt Ridley describes in “When ideas have sex” that it’s all about creating and sharing (he calls it trading), see [http://www.ted.com/talks/matt\\_ridley\\_when\\_ideas\\_have\\_sex.html](http://www.ted.com/talks/matt_ridley_when_ideas_have_sex.html).

sciences, economics, and more. Much can be learned by viewing a network-based situation from an alternative disciplinary perspective.

A crucial issue in this context is grasping the dynamics of networks at a conceptual as well as a methodological level:

- Levers of change include technology, as it has proliferated across societies;
- spill-over effects across domains, e.g., the public, political, and commercial domain;
- counter-forces, dark networks show a similar dynamics;
- innovation and defective behavior: innovation is often driven by defective behavior, e.g. young people challenging the power of global media companies;
- methodologically, e.g., living labs.

Science is currently driven by the fast development and changing character of social networks. Taking into account the high relevance of understanding the dynamics of networks, only an inter-disciplinary view on the different aspects of networks could develop the chance to grasp the nature of networks dynamics. A methodological mesh of different approaches used in the various disciplines could be a promising way to tackle the numerous research questions. Furthermore the comparison of the different network characteristics (social, business, logistics, etc.) and investigating the possibilities of transferring principles between the different network types could bring up new ways of understanding and managing these networks.

Business networks will grow (we will see more and different shaped business networks), personal networks will change (Facebook in the future will not be as Facebook is right now). We will see an integration of business networks and personal networks (social networking platforms like Facebook, XING, LinkedIn etc. will be integrated parts of businesses and business networks). Law will not be able to cover all the implications of computer-supported networks and will lose its controlling function. Interdisciplinary research is necessary to recognize, to describe, to explain and, even more important, to design and to innovate social, supply-chain, administrative, business, commerce, and political networks.

We are stuck in the silos of our disciplines. One crucial aspect that will help to change this situation is student training. Key takeaways are:

1. We need to get ahead of the curve, i.e., instead of dealing with old networks, we need to understand how to monitor dynamically and predict the development of current and future networks at some appropriate level of abstraction.
2. The notion that methods working for offline networks (viz. process mining in supply chains and/or enterprises) can be used for exploring online mechanisms needs to be explored more.
3. The traditional economic models applied for networks may be from a wrong perspective; for example, the increasing interest (by commercial providers such as Factual, Socrata, or Kasabi, to name just a few) to view data as “goods” that have a price tag and that can be traded on a market may help to explain changes and predict needs for a Web free of data issues.

In particular, we should strive to make methods of IS research, such as business process management, decision support systems or data mining, better accessible for investigating phenomena of social networking.

The most important observation is that networks cross all disciplinary boundaries. Research communities are discovering this at the moment, while funding agencies are not. Indeed, it is important to convince funding agencies that multi-disciplinary research should arrive on their agendas. Moreover, Web sciences need to be developed as a field, and also need to be integrated into teaching. This will most likely lead to novel curricula which receive their content from multiple disciplines in a balanced way.

We are at the dawn of a new way of doing research, namely detached from the fact that “I belong to a particular department”, “I need to publish in certain journals”, “I get funding only in my field”. Instead, we are overcoming field boundaries and diving into other areas together with new people. That applies even to the workshop acceptance criteria at Dagstuhl itself. We need a problem-oriented approach to get away from silo thinking: What is the problem? What expertise is needed to solve it?

Web-based systems will transform society: large numbers of users can interact; the available technology enables communities to build and run their own social machines. For a platform to be successful, it should not crack or allow for a bad experience, which requires more than a research prototype. Instead, we need professional software developers. We also need advertising, which is not funded either. The emerging area of *Web Science*<sup>8</sup> has apparently recognized this need, and is working in various ways on bridging the gaps between disciplines [9].

Intuitions we have on certain aspects seem to be wrong most of the time; this is what we see in the network domain (e.g. growth of Facebook). Now that fields are starting to converge, this is even more true. Therefore we need to start looking into the real problem. Economics enjoy modeling, but at the price of complexity reduction. In the micro/meso/macro layer setting, we need to analyze the dynamics between these layers. Studying dynamics becomes even harder that way. A good example is human-computer interaction (HCI). Interfaces were studied for years because they stood still during that time; that is not the case anymore. Now you have to know your users in advance when you build a system. HCI has developed methodologies, which have broken down. This example is about speed of change, not networking. Also in other examples, speed of change is a big issue.

## Acknowledgement

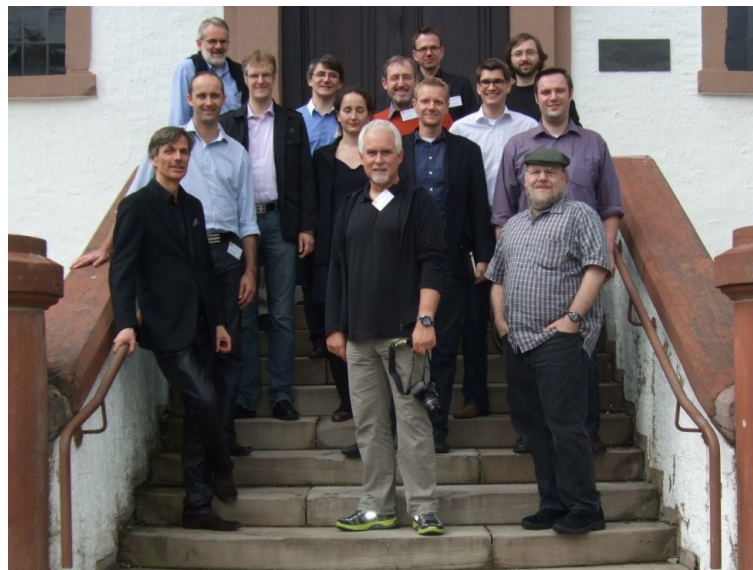
We thank the participants of Dagstuhl Perspectives Workshop 12182 for their valuable contributions.

---

<sup>8</sup> <http://webscience.org>  
<http://eprints.soton.ac.uk/265186/1/metadataisthemessage.pdf>  
<http://journals.cambridge.org/action/displaySpecialPage?pageId=3656>

## 7 Participants

- Jörg Becker  
Universität Münster, DE
- Daniel Beverungen  
Universität Münster, DE
- François Bry  
LMU München, DE
- Clemens Cap  
Universität Rostock, DE
- Ingo Dahm  
Deutsche Telekom – Bonn, DE
- Stuart Dillon  
University of Waikato, NZ
- Emese Domahidi  
Universität Hohenheim, DE
- Matthias Häsel  
Otto Group – Hamburg, DE
- Jim Hendler  
Rensselaer Polytechnic, US
- Bernd Hellingrath  
Universität Münster, DE
- Stefan Klein  
Universität Münster, DE
- Nicolas Pflanzl  
Universität Münster, DE
- Thorsten Quandt  
Universität Hohenheim, DE
- Michael Räckers  
Universität Münster, DE
- Gottfried Vossen  
Universität Münster, DE




---

## References

- 1 Enrique Estellés Arolas and Fernando González-Ladrón de Guevara. Towards an integrated crowdsourcing definition. *J. Information Science*, 38(2):189–200, 2012.
- 2 Tim Berners-Lee and Mark Fischetti. *Weaving the web – the original design and ultimate destiny of the World Wide Web by its inventor*. HarperBusiness, 2000.
- 3 Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data – the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
- 4 Peter Buneman, James Cheney, Sam Lindley, and Heiko Müller. Dbwiki: a structured wiki for curated data and collaborative data management. In Sellis et al. [11], pages 1335–1338.
- 5 Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and where: A characterization of data provenance. In Jan Van den Bussche and Victor Vianu, editors, *ICDT*, volume 1973 of *Lecture Notes in Computer Science*, pages 316–330. Springer, 2001.
- 6 Peter Buneman and Wang Chiew Tan. Provenance in databases. In Chee Yong Chan, Beng Chin Ooi, and Aoying Zhou, editors, *SIGMOD Conference*, pages 1171–1173. ACM, 2007.
- 7 Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. Crowddb: answering queries with crowdsourcing. In Sellis et al. [11], pages 61–72.

- 8 Matthias Häsel. Opensocial: an enabler for social applications on the web. *Commun. ACM*, 54(1):139–144, January 2011.
- 9 James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, July 2008.
- 10 Stefano Rizzi. Data warehouse. In Benjamin W. Wah, editor, *Wiley Encyclopedia of Computer Science and Engineering*. Wiley, 2008.
- 11 Timos K. Sellis, Renée J. Miller, Anastasios Kementsietsidis, and Yannis Velegarakis, editors. *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*. ACM, 2011.
- 12 Florian Stahl, Fabian Schomm, and Gottfried Vossen. Marketplaces for data: An initial survey. *Working Paper No. 12, European Research Center for Information Systems, Münster, Germany*, 2012.
- 13 Gottfried Vossen and Stephan Hagemann. *Unleashing Web 2.0: From Concepts to Creativity*. Morgan Kaufman, 2007.

# Computation and Palaeography: Potentials and Limits\*

Edited by

Tal Hassner<sup>1</sup>, Malte Rehbein<sup>2</sup>, Peter A. Stokes<sup>3</sup>, and Lior Wolf<sup>4</sup>

- 1 Department of Mathematics and Computer Science, The Open University of Israel, IL, [hassner@openu.ac.il](mailto:hassner@openu.ac.il)
- 2 Department of History, University of Nebraska – Lincoln, USA, [malte.rehbein@unl.edu](mailto:malte.rehbein@unl.edu)
- 3 Department of Digital Humanities, King’s College London, UK, [peter.stokes@kcl.ac.uk](mailto:peter.stokes@kcl.ac.uk)
- 4 The Blavatnik School of Computer Science, Tel Aviv University, IL, [wolf@cs.tau.ac.il](mailto:wolf@cs.tau.ac.il)

---

## Abstract

This manifesto 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 and labor consuming 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.

**Perspectives Workshop** 18.–21. Sept., 2012 – [www.dagstuhl.de/12382](http://www.dagstuhl.de/12382)

**1998 ACM Subject Classification** I.5.4 Applications (Text processing, Computer vision), I.7 Document and Text Processing, H.3.7 Digital Libraries, J.5 Arts and Humanities (Literature)

**Keywords and phrases** Digital Humanities, Digital Palaeography, Cultural Heritage

**Digital Object Identifier** 10.4230/DagMan.2.1.14

## Executive Summary

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 (a full report of which is available in [18]). 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,

---

\* Author names in alphabetical order



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

Computation and Palaeography: Potentials and Limits, *Dagstuhl Manifestos*, Vol. 2, Issue 1, pp. 14–35  
Editors: T. Hassner, M. Rehbein, P.A. Stokes, and L. Wolf



DAGSTUHL  
MANIFESTOS

Dagstuhl Manifestos  
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 palaeographer 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.

In addition to these themes, several specific recommendations regarding research infrastructure and support were made. These include:

1. A clear articulation of standards for digital image acquisition followed by all digital imaging projects when possible.
2. EU-wide harmonisation of copyright and licensing practices. 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.
3. Ideally, set copyright appropriately to allow for 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.
4. Encouraging an interdisciplinary research agenda including disciplines dealing with computable images from various perspectives such as medical imaging, cognitive sciences, Cultural Heritage Imaging (CHI), or Natural Language Processing (NLP).

This manifesto elaborates on the existing challenges and limitations of the field and details the long-term recommendations that have emerged in the workshop.

 **Table of Contents**

<b>Executive Summary</b> . . . . .	14
<b>Introduction</b> . . . . .	17
<b>Computation in Palaeography</b> . . . . .	18
State-of-the-Art . . . . .	18
Challenges . . . . .	19
Needs . . . . .	20
<b>Towards a Research Agenda for Computation and Palaeography</b> . . . . .	23
Challenges . . . . .	23
Directions . . . . .	26
<b>Participants</b> . . . . .	32
<b>References</b> . . . . .	32

## 1 Introduction

Manuscripts are the most important witnesses to and artefacts from our shared cultural heritage of the European Middle Ages. Current estimates are that close to one million manuscript books survive along with countless archival documents from a period stretching across more than a millennium. Cumulatively, these documents are the chief sources of history, history of science, literature, and art history (due to the presence of manuscript decoration) from that period. Moreover, these manuscripts are important subjects of scientific enquiry in their own right, as they bear witness to the history of the book, to scribal and monastic culture, the history of the development of handwriting systems, languages and dialects, the history and genealogy of texts over time, and the evolution of strategies for organizing texts and knowledge.

Although often taken more broadly, palaeography is in essence the study of old handwriting from manuscripts. As such, palaeographers are often asked one of four questions regarding manuscript documents from the past: what was written, when was this written, where was it written and by whom. Answering these questions, and indeed reading the text itself, are basic prerequisites for any kind of work with primary sources, and the study of almost all fields relevant to the ancient and medieval past therefore depends on them. In this respect palaeography is sometimes regarded as a “mere” auxiliary discipline. However, palaeography also extends beyond this: it encompasses the history of one of humanity’s most pervasive technologies – writing – and therefore raises questions of cultural history, the development and spread of ideas, and so on, along with the deep understanding of the transmission and use of texts which it brings. Misunderstandings here can lead to significant errors in scholarship, such as basing historical arguments on charters which prove to be late forgeries [47], or conducting studies of spelling and automatic authorship attribution without considering the effects of textual transmission, both scribal and editorial, and the changes that this brings [49].

Palaeography as a discipline typically involves difficult, complex, and time-consuming tasks, often involving reference to a variety of linguistic and archaeological data sets, and the invocation of previous knowledge of similar documentary material. Due to the involved reading process, it is difficult to record how the final interpretation of the document was reached, and which competing hypotheses were presented, adopted, or discarded in the process. It is also difficult to acknowledge and present the probabilities and uncertainties which were called on to resolve a final reading of a text. As a result, palaeographical discussion tends towards assertions based on experience with little supporting evidence – sometimes none at all – and this has led to an allegedly “authoritarian” discipline which depends on “faith” [10] or “dogma” [16] and is based on “informed guesswork” [16]. It is perhaps no surprise that the discipline itself suffers as a result [10, 3].

Palaeography as a discipline is, however, of high relevance for society and economy. All of the world’s written heritage was written by hand until the invention of printing, and texts written by hand have remained important ever since. Manuscripts are hence one of the major sources of knowledge of human culture and society, crossing the borders of modern nations, for most of what we call history. However, unlike printed texts which are distributed through libraries, handwritten sources are often accessible only to a very small and highly trained group of experts, and hundreds of thousands if not millions of manuscripts are scattered around the world. They can be difficult to find and difficult to read, are often written in an old language, and frequently deal with a subject matter that can be understood only by experts. On the other hand, however, they can be a valuable resource also for

public interest such as regional economies and tourism, as demonstrated by highly successful exhibitions which charge for entry such as the Book of Kells at Trinity College in Dublin, a book which was also an inspiration for creativity and the generation of further derived art. There are relatively few examples of manuscripts exploited in this way, but this material remains important for connecting people with their heritage and fostering identity, be it local, regional, national or pan-national.

Research can enhance and popularize the access to this largely untapped resource and can increase the number of beneficiaries of the documents. This is an investment that may bring large returns in the long term. In addition, the area of digital palaeography which is examined in this manifesto promotes technical research in challenging problems, such as processing of ancient documents, and can help develop techniques that may be helpful in other areas.

## **2** Computation in Palaeography

### **2.1** State-of-the-Art

Partly in response to the perception of palaeography as “dogma”, scholars worldwide have been developing and employing new technologies and computer-based methods for palaeographic research. This approach, often referred to as Digital Palaeography [7] and situated in the wider field of Digital Humanities, aims to improve and enhance the traditional methods. Its goal is to help efficiently solve palaeographic issues and/or provide more quantitative evidence to palaeographical arguments, and in consequence to cater for a better understanding of our cultural heritage.

As of today, there are numerous projects concerned with developing such methodologies. These encompass a wide range of scientific, interdisciplinary approaches such as forensic document analysis, optical character recognition, quantification of “scribal fingerprints”, metric analysis, quantitative methods, advanced manuscript analyses such as DNA and imaging techniques such as multi-spectral digitisation, classification systems and databases. Although some achievements have been made already, much research is still required. For instance, something as seemingly fundamental as the automated recognition of characters in handwritten texts has proven extremely complex, due largely to the very wide variation in styles of handwriting, the often poor quality of surviving manuscripts, the lack of standard orthographies which complicates prediction, and so on.

Such computational methods as proposed by digital palaeography have been the subject of research in the last few years, but most of this has been theoretical or applied only to small cases, partly because of the very high degree of labour that is typically involved [7, 2, 48, 40]. The applications to date have also focused almost exclusively on the question of scribal identity, ignoring other aspects of palaeography. Furthermore, they tend to view letter-forms as objects outside the manuscript or documentary context in which they were written, but palaeographers have long understood that handwriting depends heavily on the context in which it is produced ([4, 53], among many). Much more significantly, these methods tend to make the computer a “black box” which receives images of manuscripts at one end and returns a classification of the handwriting at the other (for examples see [40]). However, they are normally heavily dependent on very subtle and often unstated assumptions about the underlying data [44], but it is difficult or impossible for “traditional” palaeographers to evaluate these, so that usually scholars cannot evaluate the “black box” and so are

rightly reluctant to accept its results [49, 58, 9, 43]. The major challenge for computational approaches is to provide a system which presents palaeographical data quickly and easily in a way which scholars can understand, evaluate, and trust. The success and impact of research and initiatives in computational methods so far ([54] with publications [40, 15, 13, 30, 5]) has shown the strong need to combine scientific computing and palaeography in order to further investigate the interdisciplinary methods and scientific fields. It is also apparent that no institution – let alone a single scholar – is capable of undertaking comprehensive research that encompasses all those methods (and potentially more). Thus, a joint effort is required, preferably on an international level.

As became very evident during scholarly meetings on this topic [54, 62, 5], palaeographers and computer scientists tend to think in different terms and tend not to agree even on very basic notions such as “evidence” or “meaning.” Successful collaboration between researchers in humanities and in computer science is not nearly as simple as “define a computational problem and find an algorithm to solve it.” The input is often loosely defined, and the output needs to be more than just a score on some abstract scale. It is therefore crucial to identify a common level at which effective communication can be established.

## 2.2 Challenges

During the Dagstuhl Perspectives Workshop, the unmediated interaction between palaeographers and computer scientists yielded several very clear questions and themes for the future of research in Digital Palaeography. These include the following four challenges:

1. *How to optimize collaboration between all the different domain experts involved in Digital Palaeography.*

Barriers in communication between palaeographers and computer scientists are a prevailing problem. This is often reflected not only in computerized tools failing to meet the requirements of palaeographers but also in the different terminologies used by the two disciplines. It was recommended that better communication should be fostered by joint events and long-term collaborations.

2. *How to ensure that palaeographers remain in control of their research, whilst taking advantage of the possibilities of computerized approaches.*

Computerized palaeographic tools are often “black boxes” putting palaeographers on one end of the system, only receiving a system’s output, with little opportunity to directly influence how the system performs, or to communicate with it by using natural palaeographic terminology. The long-term desire is to have the scholar at the centre of the computerized system, allowing interaction and feedback in order both to fine-tune performance and to interpret and communicate results more effectively. 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.

3. *How to facilitate sharing, not only of palaeographical data and results, but also of the methodologies involved in palaeography generally.*

Palaeographic data is scarce and access to it is restricted by copyright and indexing problems, non-standard benchmarking techniques, and the lack of accurate meta-data and ground-truth information. During the workshop, multiple opportunities were identified for the acquisition of data and for its dissemination in the palaeographic research community and to the wider public.

4. *How to use the outreach potential offered by computerized technologies to enrich palaeographical knowledge.*

Palaeographic research is an expert domain. 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 other opportunities, such as crowd sourcing and others which have proved successful in other domains.

## 2.3 Needs

In this manifesto we address both the technical aspects of the collaboration between computer scientists and humanists as well as conceptual tools such as “mid level features” and “ontologies” (discussed below) that can serve as means for effective communication among practitioners. The emphasis of this discussion is not on the most efficient algorithm, producing the most accurate results. It is also not on the least ambiguous and most meaningful definitions. Instead, the emphasis is on the most effective and fruitful communication.

### 2.3.1 Data Acquisition

Repositories across the European Union have been engaged in large-scale digitization efforts in recent years, resulting in collections of hundreds of thousands or even millions of digital images of manuscript books and materials. Digital Palaeography relies on the existence of these digital surrogates of manuscripts. Moreover, some of the most exciting prospects of this field can only be demonstrated on sufficiently large collections. However, enabling this first requires modification of both policies and acquisition practices.

Specifically, from the computer user’s perspective, obtaining digital copies calls for suitable procedures and for standardization. Recently, Shweka et al. have suggested specific practices drawing on their experience in large-scale digitization [45]. These suggestions range from minimum resolution, to the usage of particular rulers and background, and also include suggested policies regarding availability and manipulability during viewing. It is emphasized that taking into account the potential usage of a computer system to analyze the image does not degrade the experience of the human viewer. For example, while image analysis is much easier on a blue/green background, for human viewing purposes, such a background can be easily replaced. We propose the following:

1. A clear articulation of standards for digital image acquisition followed by all digital imaging projects when possible. Where such standards already exists (e.g. “DFG-Praxisregeln ‘Digitalisierung’ ” of the German Research Council DFG [17] or JISC Guidelines in the UK [23]), they should be checked against the requirements of palaeographers (see also [56]) and, if necessary, extended to encompass and meet them in full. These include practices such as:
  - Proper use of colour bars and grey cards.
  - Appropriate use and documentation of illumination and equipment (e.g. lighting parameters including positioning, hardware).
  - References to size of original objects using shared standards.
  - Metadata descriptions of digitized objects following internationally accepted standards such as MIX/METS; if one takes several images of the same object (e.g. different lighting, multiple sizes, multispectral), it is important that the corresponding metadata

indicates that these are images of the same object, and what the relationship between the images is.

- Information that links multiple names and catalogue records when original objects have no single identifier (e.g., a manuscript with shelf marks that change over time and that is also referred to by other common names in scholarly literature).
  - 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.
2. 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.
  3. EU-wide harmonisation of copyright and licensing practices. 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 (cf. [33] and [59]).
  4. 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 [42]. 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. 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).
  5. Freedom of resources produced by cultural institutions must be actively encouraged because it benefits the owners and enables research. The more it generates connections the more it becomes valuable: as well as research connections, it also generates connections back to the institutions themselves, bringing value to those institutions (as demonstrated by examples such as [11], for which see further below).
  6. Encouraging an interdisciplinary research agenda including disciplines dealing with computable images from various perspectives such as medical imaging, cognitive sciences, Cultural Heritage Imaging (CHI), or Natural Language Processing (NLP).

### 2.3.2 Tools, Libraries and Resources

The overall objective of tools, software libraries, and resources to be developed in the context of palaeography is to provide support in establishing the correlation between text as shape and text as meaning; which, in the most general of senses, can also be understood as one of the aims of palaeography as a subject.

The starting point is to firmly acknowledge and map out the domains of expertise of the agents involved in the process, namely, humans and computer-based tools. On the one hand, computers excel at dealing with “big data”, namely at tasks ranging from holding large amounts of data in memory to carrying out process-intensive computations such as the identification of fine differences and rare occurrences within large datasets. On the other hand, humans (including palaeographers) excel at dealing with data which is ambiguous, complex, or broad, in the sense that the datasets are made of heterogeneous pieces of data. Humans also excel at making sense of the data, at expressing its gestalt in the sense that the whole of the data expresses more than the sum of its parts.

Taking these distinct sets of skills into account, the highest priority in developing computational resources for palaeography is the production of semi-automatic and interactive

tools, where palaeographers can continually intervene, inform, correct, understand, use, and reuse results produced by and processes implemented by these tools. Only in this manner will palaeography benefit optimally from the respective strengths of the human and computational agents. Ideally, developing such semi-automatic and interactive tools will stimulate the establishment of a mutually beneficial continuous feedback loop between human and machine, whereby humans will be involved at all levels of reasoning, machines will be able to learn from human input, and palaeographers and others will learn and create new knowledge more effectively through the use of machines [22, 25].

We recognise that a critical mass of data is required for performing research, and the pre-attentive perception of the data by researchers is a major factor in building new hypotheses. This critical mass of data can, on the other hand, only be obtained through usable and ergonomic tools. Hence, in tool development for Digital Palaeography, focus groups, user testing and proper user interface design is needed in consultation with humanities scholars as end-users (for which see also [24] and [26]). A further requirement that emerges from here is the recognition of tool-development as academic research to encourage Digital Humanities scholars to publish their work and make it usable by a broader audience.

In the following, we outline the specific levels at which helpful computational tools can be developed as well as possible ways of keeping the humans in the loop. All tools developed should be compatible with one another and combinable at will (or, more precisely, as long as the notions involved are compatible, the tools should be). They might be used sequentially, or contribute to one another. We have identified the following categories of multi-level computational tools for Digital Palaeography:

1. Low-level tools:
  - Binarization
  - Segmentation
  - Alignment, matching and registration of features (for similarity measures) including expert features of handwriting extraction (e.g. angles, curvatures, strokes)
  - Physical feature extraction
  - Similarity measures (for comparison between characters, words, texts, fragments, documents, corpora)
2. Mid-level tools:
  - Clustering
  - Classification
  - Character recognition
  - Word spotting
  - Cross-modality search engines, where the input for the searches might not be in the same form as the dataset that is searched, e.g.:
    - Search for a string in a text / corpus
    - Search for an image in a text / corpus
    - Search for a string in an image / a set of images
    - Search for an image in an image / a set of images
    - Search for a shape (shape would here be a hand-drawn input e.g. SVG, as opposed to an image that would be in a rasterised format)
  - Image-text (shape-meaning) correlation
3. Databases, where the data is organised in a way that allows fast queries of (for example):
  - Metadata
  - Transcripts
  - Images



- Properties of the text (author, genre, date etc.)
  - Scripts and scribal features
4. Higher-level tools:
- Interfaces, ergonomics, user experience (“UI”/“UX”)
  - Searches of combinations of characters/words (bigrams, trigrams, possibly of shapes and/or images)
  - Correspondences in expert vocabularies
  - Inferences of paraphrases and synonyms for searches through metadata (widening searches by applying fuzzy techniques on search terms, by proceeding by analogy, etc.)
  - Web services
  - Web-based research environments for online collaboration and benchmarking within a global community.

Approaches and tools that keep humans in the loop can further be classified along two main lines: data acquisition/exchange, and cognitive triggers/feedback loops. These include:

1. Data acquisition and exchange:
  - Provision of training data / annotated data
  - Online training / expert-in-the-loop
  - Crowd-sourcing
2. Feedback loops and cognitive triggers:
  - Drawing / touch screen technologies
  - Simple interactive image enhancements
  - Visualization aspects of interactions with all the tools listed above (of results, of databases), interactive visualisations – e.g., time varying graphs – with customizability as a priority [22, 26]
  - Rationale building support, tracking of expert hypotheses in interpretation building
  - Statistical tools – with tests of significance
  - Information sharing systems
  - Transcription tools linking text and image.

## **3** Towards a Research Agenda for Computation and Palaeography

### **3.1** Challenges

This section is focused more on challenges than on constraints. We use the term “challenge” because it seems that, although the hurdles presented below do constitute some forms of limitations, we do not believe them to be insurmountable.

#### **3.1.1** Context and Meaning

The first observation is that something is generally excluded from systematic analysis, namely the interpretation of data. Contextual knowledge and meaning, which are required for interpretation, are both concepts that are usually best handled by humans because they involve unstructured and non-formalised knowledge. This type of knowledge is often implicit in the natural scholarly environment, and although attempts can be made to structure and formalise contextual knowledge and sense-making processes, their continual evolution dooms the task to permanent incompleteness – which justifies the palaeographers’ wish for more involvement and interactivity at various levels of computational formulation and formalization.

### 3.1.2 Access to Data

The second observation is that, beyond context and meaning, current computational constraints are usually related either to access to data or to data retrieval. The problem of access to data is largely political in nature, therefore differing between countries and bylaws involved (e.g. in the UK, access and use of images can be drastically restricted due to copyright and licensing issues: see “Data Acquisition” above). In case of data retrieval, it is largely the degree of flexibility of the search tools that limits their usability, their usefulness and thereby their use. Search tools often present difficulties of precision and recall, and this is usually due to parameters being either too inflexible or, paradoxically, too flexible. This mismatch between the flexibility needed and the flexibility provided by the search tools is in fact a good illustration of what we have identified as the major bottleneck in the collaboration between computational and palaeographical research, and that bottleneck once again is down to communication.

It might seem at first that problems in communication are easy to solve, and that it is “just” a matter of listening and understanding, a matter of ironing out differences. However, even in our group of twenty people at Dagstuhl from different backgrounds, where all were accustomed to collaborative scholarship, a striking recurring difficulty in understanding each other was apparent – a trait that would most definitely be accentuated in a larger group and in a group where cross-disciplinary communication is not a current practice. Some of the symptoms of this problem are enumerated below, as well as some examples, and possible measures to treat them. Unfortunately, these hints for solutions will only treat the symptoms; the roots of the problem run much deeper and are mostly cultural, originating in the traditions of each and every discipline.

### 3.1.3 Interdisciplinary Approaches to Research

Scholarly endeavours are all rooted in their own traditions. In spite of our non-subscription to Snow’s Two Cultures theory [46] – the next section on terminology will illustrate how the business of creating knowledge is a Many-Cultures system rather than a Two-Cultures system – one of the high-level observations about research methodologies in palaeography and in computer science is that they differ widely. As noted above, computer scientists tend to be problem solvers. Their approach to knowledge creation is typically to break down a large task into smaller tasks and then to solve these tasks, iteratively, until a satisfactory solution of the initial large task is found (where “satisfactory” is often left to their own discretion). In the tradition of computer sciences, there is a further convention of not deriving natural interpretation from the methodology. In other words, the output needs additional cognitive processing to be interpreted, and computer sciences do not traditionally have ways of doing so. In contrast, palaeographers tend to approach knowledge creation in a different way. Their method is typically to derive questions from questions, where a new question often has the value of answering the preceding question (see further “Exploring and Questioning, not Answering”, below).

Communicating between these two approaches can evoke situations in which what may first seem to be a misunderstanding or misconception turns out, finally, to release synergies. Take, for instance, a question in palaeography for which a corresponding computational solution has been developed. The discussion between palaeographers and computer scientists might lead to an emphasis on the weaknesses or incompleteness of the proposed computational solution. But it might also reveal the need to reformulate the original question, or might open up the potential for new, related questions. In any case, this interdisciplinary communication

helps to augment research on either side, and ideally on both sides.

It seems, therefore, that practices exist by which scholars operate at different levels of abstraction and explicitness; for example, palaeographers' relatively abstract way of formulating problems might not translate well into formal computer language. Being aware of these different modes of communication might help to smooth out some of the difficulties and minimize possible frustrations, but the differences in traditions are not likely to change much, and the problems of terminology remain to be addressed. Indeed, these differences are strengths insofar as they allow approaches to different types of questions, and so they should be embraced rather than ignored or suppressed.

### 3.1.4 Terminology

As hinted above, the differences in research cultures are deeper than different methodological approaches to research (e.g. questioning versus problem-solving). For example, the use of specialized terminology in each domain, where words can coincide but carry different meanings, presents a much greater challenge than is apparent at first sight. A telling example is that of the word "feature". In image processing, "feature" has a very specific meaning: it describes a defined behaviour in terms of signal, an idealised profile such as a step, a ridge, a trough. In palaeography, too, the word "feature" is used with a very specific meaning; it describes the aspects of a stroke that make it characteristic of a certain hand, a certain scribal school, a certain area, or a certain type of document (e.g. its ductus, or the variation in its width). The two domains have therefore their own typical – i.e. accepted and shared – use for the word within their community, but this usage does not translate smoothly from one community to the other. This example is only one of the many that illustrate the terminological challenges that might be encountered (some others are "ontology" and "pattern" which are discussed further below).

It is also worth noting that this issue with the uses of specific terms in various disciplines constitutes a bottleneck in communication not only between computer scientists and palaeographers. Within the computer sciences themselves, communities such the data mining community and the image processing community also share some words, but not necessarily the meaning attached to them ("feature" is an example once again). Similarly, palaeography has long been troubled by differences in terminology, despite the best efforts of the Comité international de paléographie latine and others to standardise them. The differences run deeper than simple choice of words: expert vocabularies in each discipline and in each sub-domain carry their own implicit contexts and assumptions that can prevent people from understanding each other across and even within fields (cf. [38] and [10]).

### 3.1.5 The Problem of the Black Box

The last type of bottleneck for communication and mutual understanding across scholarly disciplines resides in the fact that expertise implies tacit knowledge, and tacit knowledge tends to produce "black boxes", namely systems – whether human or machine – which take inputs and produce results without giving any indication of how those results were obtained. Computational algorithms are often perceived as black boxes by palaeographers, and palaeographical expertise is also seen as a kind of black box by computer scientists and indeed by other experts in the Humanities. The main issue here is to not concentrate exclusively on "cracking open" the black boxes to understand all the internal nuts and bolts that power them. Rather what is required is the establishment of trust between the communities. This trust might best be created by communicating an understanding of the

principles and assumptions behind the inner working of the black boxes and not of the details of the methods and their implementation. Establishing that trust will alleviate the anxieties that black boxes tend to generate; it will thereby ease communication and collaboration.

Two (non-exclusive) natural solutions to such bottlenecks and lack of trust can be summarized as the introduction of an “in-betweener” and communication of “mid-level features”; both of these are discussed further below.

### 3.2 Directions

It is worth noting that the technical limitations outlined above are not reviewed in more detail here because, in the light of the potential problems in communication already discussed, they seem largely surmountable. In fact, through the discussions, round tables, and Q&A-sessions during the Dagstuhl workshop, it often emerged that computational approaches offer a lot more possibilities than single experts might have predicted. As a result, any prognosis of technical limitations voiced here would carry the inherent risk of outlining pre-emptive delimitations.

#### 3.2.1 Interdisciplinarity and the “In-Betweener”

The Dagstuhl workshop can serve as a best-practice or “template” for future interdisciplinary communication. Further joint sessions at conferences and similar events need to be held. But communication between computer scientists on the one hand and Humanities scholars on the other is only a starting point. Interdisciplinary projects between the fields need to be strengthened, and all participating disciplines will draw significant benefits from them. Experts in scientific computing should not merely implement requirements formulated by the Humanities, but should also suggest ideas based on their excellence and expertise. At the same time, scholars in Computer Science should acknowledge the relevance of research questions and methods from the Humanities. Although the disciplines have different semiotics and separate proof systems, interdisciplinary communication and cooperation leads to better understanding and consequently to new knowledge.

Interdisciplinary workshops are invaluable, but also necessary is the interdisciplinary individual: the “in-betweener” introduced above. This is a middle-person, a translator: a person who is versed enough in each of the collaborating fields to understand enough of each of the discipline-specific lexical fields to foster good communication and fruitful exchanges. Dedicated specialized “in-betweeners” have already been used very successfully in some Digital Humanities contexts, such as the positions of “project analysts” at the Department of Digital Humanities in King’s College London, and their application to palaeography is to be encouraged.

#### 3.2.2 Communication, Intelligibility and the “Black Box”: Evidence-based Palaeography

Given the task of classifying a written fragment, an authoritative palaeographer might examine the page and simply state his or her classification of it, typically providing little evidence for how this conclusion was reached [16, 10, 9]. Somewhat analogously, given an image of a fragment, a computerized system might output the class of script for which the fragment scored the highest, along with the score itself. The mathematical procedures and formulas that led to this conclusion would remain inaccessible inside the “black boxes.” Both

the authoritative palaeographer and the computer leave little room for further discussion or debate on the results, and their work is therefore somewhat limited in expanding the science of palaeography, even though the answer might still be of a great help to a historian, for example, working on that specific manuscript.

Here, we suggest establishing a system for palaeographic representation which is accessible to both scholars and machines and can serve as the foundation of an evidence-based palaeography.

This representation system would rely on “mid-level” features or descriptors as introduced above. The mid-level features seek to define a shared vocabulary between disciplines, a shared meeting ground where each field can intervene with its own perspective. The term “mid-level” here means that these descriptors require visual identification, unlike low-level features that are extracted computationally from the images and which cannot readily be verified by a human observer. This identification is meant to be as unambiguous as possible, such that if one researcher or computer system identifies or detects that such a feature exists in the test, other researchers or systems can verify this claim. In other words, one should be able to dispute almost completely on a factual basis any evidence that is structured according to these mid-level features.

Determining the mid-level qualifier is crucial: communication needs to be more fine-grained than any abstract conceptual discussion around principles would be; and it must not become bogged down in the (sometimes murky) fine details. Specifically, these features must hold high-level meanings to the palaeographers on one hand, but must still be concrete enough to be definable in terms of a computerized system on the other. From a computational point of view, this is similar to the way by which facial features are used to identify faces in photos [61]. An example candidate for mid-level features are palaeographic “letter features”, used to describe and identify handwriting. These are amenable to computerized analysis [29]. This approach has the inherent risk of systematising and formulating the field-specific strategies, thereby possibly compromising the potential for creativity as well as the integrity of each discipline, slowing down progress and over-constraining the problem spaces. However, this seems to be a more than acceptable compromise compared to the risks carried by repeated breaks in communication and failed exchanges. The approach therefore warrants much more investigation, and as a starting-point mid-level features and their application to the “black box” problem are discussed further below.

The use of mid-level features requires both the authoritative palaeographers and some of the computer systems to adapt. Palaeographers, like experts in other domains, often cannot provide the rationale that led to their decisions. In data analysis, systems that are built for maximal accuracy are built to solve the specific task at hand, and not necessarily to rely on rules that are interpretable; adding to the requirement of accuracy the requirement of interpretability would typically hurt the performance of the system. Despite these adaptations, the potential benefits are very significant. While scholars are reluctant to use the output of black-box systems, we expect much easier adaptation to computer systems which provide clear evidence for their classification. As was discussed during the Dagstuhl Workshop, the choice is between having an accurate system that lies unused and having a somewhat less accurate system that scholars are happy to employ.

### 3.2.3 Ontologies instead of Terminologies

Difficulties in communication have arisen several times in the discussion already, including not only communication between disciplines but even within them. In particular, many efforts have been devoted to creating a unified terminology in palaeography [10]. These

efforts have met with great challenges and were not able to achieve their goals. We suggest embracing the differences in terminologies and the complex relations that exist between them and focussing instead on developing an ontology.

An ontology, in this understanding, is a representation of a knowledge domain which is based upon well defined entities, each having a unique meaning. Various structural links are then used to define relations, such as “subst of” (also known as “is a”), “related to”, and so on. Each term can also contain a list of synonyms and translations, a definition, references to other terminologies, and remarks. Instead of traditional classification systems, ontologies are being used more and more widely in Humanities scholarship, especially in cultural heritage documentation, because of the much greater flexibility that they allow. Examples of ontologies that are widely used in the Humanities include the EDM model of the Europeana library of digital objects [14]; and the CIDOC Conceptual Reference Model [20], which has become an international standard [21].

An example of part of the envisioned ontology for palaeography could be:

```
[TERM]
id: PCO0000345
name: triangular ascender
def: triangular decoration at the tip of an ascender
synonym: wedged ascender EXACT
related to:
is_a: PCO0000221 ascender decoration
```

The top level of the ontology could contain the terms “Manuscript Studies” and underneath “Palaeography” and “Codicology”. The latter could be based on an effort recently envisioned ([28], and compare also [31]). We note that the field of palaeography is much less ordered than codicology, and this can lead to challenges in representing it. Nevertheless, the top distinctions under palaeography could be “Allographs” and “Graphemes”, or something similar. Some of the terms would be descendants of terms from both these branches, e.g., “Caroline **a**” (cf. [51]).

Ontologies allow for unified treatment of metadata associated with documents as well as to mining of such resources. For example, projects like the “Medieval Electronic Scholarly Alliance” [34] and the “Manuscripts Online” project [32] aim to provide federated searches which span multiple resources. However, each resource might use a different name for exactly the same term, or the same name for different things. For example, English vernacular script of the eleventh century has been labelled “Caroline minuscule” [27], “Anglo-Saxon Round minuscule” [6], “Anglo-Saxon Vernacular minuscule” [12], and “English Vernacular minuscule” [52]. Given an ontology, it is fairly straightforward to come up with reasonable methods to expand the search to include all these terms, and then rank the combined results together, and this is an approach which those projects will use, although its application in practice is far from trivial.

In the past, some projects on building ontologies (in general) were less successful than others. One of the authors of this manifesto (TH) has participated, as a student, in the construction of an ontology for representing 3D shapes. He describes a very frustrating process in which committee members debated many minute details and which ultimately led to a tool that no researcher uses. In contrast, another author (LW) is an avid user of the Human-Phenotype-Ontology [19], which is used by clinical geneticists to describe phenotypes, many of which are visual. He has witnessed the power of the ontology in facilitating the

merging of disperse terminologies and the extremely useful data mining and classification tools that that it has entailed. Interestingly, this ontology has originated from a collection of medical data called OMIM, which was written by thousands of authors, each using their own terminology [36].

Based on discussing these cases we came to the conclusion that ontologies should rely at first on the expertise of specific authoritative palaeographers rather than on the community at large, and that they must be accompanied by datasets and computational tools that employ the ontology in question.

### 3.2.4 Exploring and Questioning, not Answering

It is increasingly being recognised in related fields of Digital Humanities that the “black and white” response often given by computational methods is incompatible with the approaches and interests of Humanities scholars. Furthermore, as already discussed here, it is very difficult to move from computational results to “real world” meaning, but for most Humanities scholars this “real world” meaning is the principal or only point of interest. It is therefore becoming increasingly evident that palaeographers prefer to harness computational methods not to provide answers to real-world questions, but rather to manage large amounts of data in ways that allow them to draw their own conclusions. Furthermore, it has been observed that cognitive processes in academic research can be enhanced through visualisation, particularly when applied to material which is inherently visual such as palaeography and manuscript studies [22, 50].

Some research questions of historical content have results which Humanities scholars can verify with a relatively high degree of confidence. One example is joins, that is, identifying pages or fragments of pages from now dismembered books. For problems like these, computational methods can usefully propose “real-world” answers, for example by providing a set of images of pages which are likely to be from the same book, and which the scholar can then check. In contrast, other problems are difficult or impossible to verify against the historical “truth,” and computational methods which attempt to answer these have not been accepted because of this difficulty of verification. Here, Humanities scholars need to be able to “cross-examine” the results, including also the method and the assumptions which underlie them [9, 43]: if they cannot verify them then they cannot have any confidence in the results. This is closely related to the “black box” problem discussed above and, as already noted, it is a significant challenge for future work.

However, an alternative approach is rather to develop computational methods that allow researchers to manipulate and visualise the content on their own terms, and to communicate this data as evidence to a broader audience. Scholars in Digital Humanities have referred to the “virtue of automated analysis ... not [as] the ready delivery of objective truth, but instead the more profound virtue of bringing us up short, of disturbing us in our preconceptions” [44]. Therefore, once a critical mass of data has been reached, problems of this sort should be treated with computational methods designed to aid discovery, exchange, interpretation, and presentation of knowledge, not providing answers to historical or other “real world” questions ([22]; cf. also [8]). This has important implications for collaborative work with computer scientists, since it is fundamentally different from the “algorithms to solve problems” approach which is more typical of the latter’s field (see “Interdisciplinary Approaches to Research”, above).

### 3.2.5 The “80/20” Problem: Working with Imperfect Results

Given the inherently ambiguous nature of data in the Humanities, it is unrealistic to expect or even aim for perfect results in the medium term. However, “imperfect” or incomplete results can still be of significant value, as there is a bottleneck of turning digitized manuscripts into texts which can be processed by a machine. Rather than attempting to “solve” this problem in the short term, further consideration is encouraged as to what can be done with computational results that are accurate to (for example) 80%, 60% and so on. Current success rates for handwriting recognition are still extremely low (as low as 30%), however, and research that promises to increase that rate should be encouraged and funded. A success rate of 80% text recognition is still bad (every fifth word would contain an error), but if it is clear which 20% are inaccurate, the 80% of data becomes usable, and following the Pareto phenomena [37], achieving these 80% becomes cheaper than focussing on the expensive remaining 20%. Furthermore, as just discussed, scholars in the Humanities do not typically expect or even desire a final, “correct” answer, but rather want tools to help them process large quantities of material. In circumstances like this, simply reducing the size of a search-space by 80% may be a very significant improvement.

This demands several prerequisites: first, computer scientists must have verifiable ways of establishing confidence in their results matching the “ground truth”: as discussed, this is often challenging or even impossible, but in some specific cases is generally achievable (e.g. text recognition and word spotting). Second, Humanities scholars must learn to understand the implications of the inaccuracies: a given type of inaccuracy will not be significant for some research questions but will be highly significant for others. Close consideration must also be given to the role of false positives versus false negatives: for example, if a computer is being used to reduce a search space which a human researcher then examines, false positives are probably expected and tolerable, but false negatives are not. These considerations again require close communication between the disciplines. Third, investment should be made in identifying new research which can be enabled by computational methods which are largely but not entirely accurate. This may include manually correcting the inaccuracies (which could still save substantially on research time), or in designing new research which is not affected by the types of inaccuracies. Close parallels already exist in fields such as computational linguistics, distant reading, and “big data” research, and lessons learned there can also be applied here.

### 3.2.6 Outreach and Dissemination

Looking beyond the academic and research audience, very significant potentials exist for outreach and dissemination of work in cultural heritage. As noted above, handwritten manuscripts and documents form a very large part of the world’s cultural heritage, with prominent examples including the Book of Kells and Lindisfarne Gospels, the Dead Sea Scrolls, through more recent examples such as Abraham Lincoln’s hand-written copy of the Gettysburg Address, or Michel Proust’s draft manuscript of *À la recherche du temps perdu* [39]. This rich cultural heritage has proven to be of great interest to a wide public, and can also help to empower minority or other disenfranchised groups and regions through informing them better of their history, heritage, language, and so on (one example of this is the Lindisfarne Gospels, which recently toured in exhibitions in North-East England). This “virtual repatriation” of cultural heritage represents a promising area of further development. More generally, however, both repositories and research institutions are frequently criticised for spending public money on material that is not accessible to those who provided the



funds, and online resources can help to overcome this. Indeed, this increased access and “democratisation” is a frequent promise of Digital Humanities, although it has not necessarily been fulfilled in practice [41].

The introduction not only of digitised images but also of computerized techniques opens up new ways of sharing this information with the broad population. One particularly effective example of this is the Walters Art Museum, whose policy of releasing digital images of manuscripts using Creative Commons licensing, and of distributing these images through a range of social and other media, has led directly to very wide public recognition of their holdings, so much so that a search for “koran” in Google Images returns a highly disproportionate number of results from that museum – far more than from much larger and better-known institutions such as the British Library or the Bibliothèque nationale de France [35]. Even more exciting is the prospect of people conducting their own research, or tapping into non-expert traditions as a way of enriching scholarly knowledge. Although the process of opening up “virtual” manuscript archives to the public has already begun, these projects are still in their infancy. Reaching out, collecting, and processing the knowledge that may be available in regional traditions, on the other hand, has not been sufficiently explored. Doing so by using “crowd sourcing” techniques is an exciting new research direction and has already been applied to transcription and identification of manuscripts and musical scores, among others (e.g. [57, 60, 1]).

In order to realise this potential fully certain requirements remain. As the Digital Walters project clearly demonstrates, one requirement is again that of sufficiently permissive copyright and licensing conditions: if people are not allowed to use images in ways that they wish, or if it is unclear whether they may so use them or not, then they typically will not use them at all [35]. The material must also be free not only of licensing restrictions but also of technical ones: again, if the images are available only in proprietary viewers or other limiting formats then access to them diminishes accordingly. Furthermore, the difficulties in communication which have been discussed between palaeographers and computer scientists become even more pronounced when moving beyond the professional researcher to the wider public. However, the same principles advocated here, such as mid-level features and “in-betweeners” specialists, are also relevant to this broader challenge. These principles need to be extended to other areas both of academic but also of public interest such as local history, genealogy, art history, language (including regional dialects), name studies, calligraphy, arts and crafts, and so on. As researchers are increasingly pressured to demonstrate the “impact” and value to society of their work, and as they discuss how best to measure and achieve it [55], digital palaeography is already addressing these concerns and also has an ideal scope of study which already has demonstrable public interest. Extending these concerns and combining the pre-existing interest presents an outstanding opportunity for taking this new and relatively marginalised field of study and bringing it to the forefront of public and academic awareness.

## Acknowledgements

The authors wish to thank Schloss Dagstuhl for their support of this workshop. PAS also thanks the European Research Council (ERC): some of the results presented here are funded by the European Union Seventh Framework Programme (FP7) under grant agreement no 263751. This text has received substantial contributions from all members of the Workshop, for which see Participants list below.

## 4 Participants

- Dimitris Arabadjis  
National TU – Athens, GR
- Nachum Dershowitz  
Tel Aviv University, IL
- Matthieu Exbrayat  
Universite d’Orleans, FR
- Shira Faigenbaum  
Tel Aviv University, IL
- Melanie Gau  
Universität Wien, AT
- Tal Hassner  
Open University – Israel, IL
- R. Manmatha  
University of Massachusetts –  
Amherst, US
- Ophir Munz-Manor  
The Open University of Israel –  
Raanan, IL
- Eyal Ofek  
Microsoft Res. – Redmond, US
- Micalis Panagopoulos  
Ionian University – Corfu, GR
- Robert Sablatnig  
Universität 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  
Ecole Pratique des Hautes  
Etudes – Paris, FR
- Segolene Tarte  
University of Oxford, GB
- Lior Wolf  
Tel Aviv University, IL




---

## References

- 1 Ancient Lives. Available online: <https://www.zooniverse.org/project/ancientlives>.
- 2 J. F. A. Aussems. *Christine de Pizan and the Scribal Fingerprint: A Quantitative Approach to Manuscript Studies*. PhD thesis, University of Utrecht, Utrecht, 2006. Available online: <http://igitur-archive.library.uu.nl/student-theses/2006-0908-200407/UUindex.html>.
- 3 Mary Beard. A Don’s Life: University cuts, redundancies – and bye-bye palaeography at King’s College London. *Times Literary Supplement*, January 28, 2010. Available online: [http://timesonline.typepad.com/dons\\_life/2010/01/university-cuts-redundancies-and-byebye-palaeography.html](http://timesonline.typepad.com/dons_life/2010/01/university-cuts-redundancies-and-byebye-palaeography.html).
- 4 Leonard E. Boyle. *Medieval Latin Palaeography: A Bibliographical Introduction*. University of Toronto Press, Toronto, 1984.
- 5 S.J. Brookes. “Digital Resources for Palaeography” One-Day Symposium. *DigiPal Project Blog*. King’s College, London, 2011. Available online: <http://www.digipal.eu/blogs/news/%E2%80%98digital-resources-for-palaeography%E2%80%99-one-day-symposium/>.
- 6 Michelle P. Brown. *A Guide to Western Historical Scripts from Antiquity to 1600*. British Library, London, 1990.
- 7 Arianna Ciula. Digital palaeography: Using the digital representation of medieval script to support palaeographic analysis. *Digital Medievalist*, 1(1), 2005. Available online: <http://www.digitalmedievalist.org/journal/1.1/ciula/>.

- 8 Tanya Clement, Sara Steger, John Unsworth, and Kirsten Uszkalo. How not to read a million books. *Seminar on the History of the Book*, Rutgers University, New Brunswick NJ, 5 March, 2009. Available online: <http://www3.isrl.uiuc.edu/~unsworth/hownot2read.rutgers.html>.
- 9 Tom Davis. The practice of handwriting identification. *The Library (7th series)*, 8(3):251–276, 2007. doi: 10.1093/library/8.3.251.
- 10 Albert Derolez. *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*. Cambridge Studies in Palaeography and Codicology, 9. Cambridge University Press, Cambridge, 2003.
- 11 Digital Walters. Baltimore: The Walters Art Museum. Available online: <http://www.thedigitalwalters.org/>.
- 12 David N. Dumville. Specimina codicum palaeoanglicorum. In *Collection of Essays in Commemoration of the 50th Anniversary of the Institute of Oriental and Occidental Studies*, pp. 1–24. Kansai University Press, Suita, Osaka, 2001.
- 13 European Science Foundation. *Humanities Exploratory Workshops: Digital Palaeography*. Available online: <http://www.esf.org/activities/exploratory-workshops/humanities-sch/workshops-detail.html?ew=10865>.
- 14 Europeana library of digital objects. *Europeana Data Model*. Available online: <http://pro.europeana.eu/edm-documentation/>.
- 15 Franz Fischer, Christiane Fritze, and Georg Vogeler, editors. *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*. Schriften des Instituts für Dokumentologie und Editorik, 3. Books on Demand, Norderstedt, 2011. Available online: <http://kups.ub.uni-koeln.de/4337/>.
- 16 David Ganz. “Editorial Palaeography”: One teacher’s suggestions. *Gazette du Livre Médiéval*, 16:17–20, 1990. Available online: <http://www.palaeographia.org/glm/glm.htm?art=ganz>.
- 17 German Research Council. *DFG-Praxisregeln. “Digitalisierung”*. 2013. Available online: [http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln\\_digitalisierung\\_2013.pdf](http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung_2013.pdf).
- 18 T. Hassner, M. Rehbein, P.A. Stokes and L. Wolf, editors. Computation and palaeography: Potentials and limits (Dagstuhl Perspectives Workshop 12382). *Dagstuhl Reports* 2:9 (2012): 184–199. doi: 10.4230/DagRep.2.9.184
- 19 HPO. *Human-Phenotype-Ontology*. Available online: <http://www.human-phenotype-ontology.org/>.
- 20 International Council of Museums - International Committee for Documentation. *The CIDOC Conceptual Reference Model*. Available online: <http://www.cidoc-crm.org/>.
- 21 International Standards Organization. *Information and Documentation – A Reference Ontology for the Interchange of Cultural Heritage Information*, 2006. Available online: [http://www.iso.org/iso/catalogue\\_detail?csnumber=34424](http://www.iso.org/iso/catalogue_detail?csnumber=34424).
- 22 Martyn Jessop. Digital visualization. *Literary and Linguistic Computing*, 23(3):281–293, 2008. doi: 10.1093/lc/fqn016
- 23 JISC. *Basic Guidelines for Image Capture and Optimisation*. Available online: <http://www.jiscdigitalmedia.ac.uk/stillimages/advice/basic-guidelines-for-image-capture-and-optimisation>.
- 24 JISC. *Usability Foundation Study and Investigation of Usability in JISC Services*, 2004. Available online: <http://www.jisc.ac.uk/whatwedo/programmes/presentation/usability.aspx>.
- 25 JISC. *Visualisation Foundation Study*, 2004. Available online: <http://www.jisc.ac.uk/whatwedo/programmes/presentation/visualization.aspx>.

- 26 JISC. *Development of Personalisation for the Information Environment 1 (DPIE1)*, 2008. Available online: <http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/dpie1.aspx>.
- 27 Neil R. Ker. *Catalogue of Manuscripts containing Anglo-Saxon*. Clarendon Press, Oxford, 1957.
- 28 Robert Kummer. Semantic technologies for manuscript descriptions-concepts and visions. In Fischer et al. [15], pp. 133–154, 2011. Available online: <http://kups.ub.uni-koeln.de/4347/>.
- 29 N. Levy, L. Wolf, N. Dershowitz, and P.A. Stokes. Estimating the distinctiveness of graphemes and allographs in palaeographic classification. In *DH2012 Book of Abstracts*, ed. by Jan Christoph Meister et al., pp. 264–267. University of Hamburg, 2012. Available online: <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/estimating-the-distinctiveness-of-graphemes-and-allographs-in-palaeographic-classification/>.
- 30 London: King’s College. *DigiPal: Digital Database and Resource of Palaeography, Manuscripts and Diplomatic*. Available online: <http://www.digipal.eu/>.
- 31 London: University of the Arts. *Ligatus: An English/Greek Terminology for the Structures and Materials of Byzantine and Greek Bookbinding*. Available online: <http://www.ligatus.org.uk/glossary/>.
- 32 *Manuscripts Online. Written Culture 1000 to 1500*. Available online: <http://www.manuscriptsonline.org/>.
- 33 Max Planck Institute. *Best Practices for Access to Images: Recommendations for Scholarly Use and Publishing*. Berlin, 2009. Available online: <http://www.mpiwg-berlin.mpg.de/PDF/MPIWGBestPracticesRecommendations.pdf>.
- 34 MESA. *Medieval Electronic Scholarly Alliance*. Available online: <http://mesa.performantsoftware.com/>.
- 35 W. Noel. *The Commons and Digital Humanities in Museums*. CUNY, New York, 2012. Available online: [http://www.youtube.com/watch?v=XPJ\\_kciC15I](http://www.youtube.com/watch?v=XPJ_kciC15I).
- 36 OMIM. *Online Mendelian Inheritance in Man*. Available online: <http://www.ncbi.nlm.nih.gov/omim/>.
- 37 Pareto. Pareto Principle. *Wikipedia*. Available online: [http://en.wikipedia.org/wiki/Pareto\\_principle](http://en.wikipedia.org/wiki/Pareto_principle).
- 38 Susannah B.F. Paletz, Christian D. Schunn, and Kevin H. Kim. The interplay of conflict and analogy in multidisciplinary teams. *Cognition*, 126(1):1–19, 2013. doi: 10.1016/j.cognition.2012.07.020.
- 39 E. Pierazzo and J. André. *Autour d’une séquence et des notes du Cahier 46: enjeu du codage dans les brouillons de Proust – Around a Sequence and some Notes of Notebook 46: Encoding Issues about Proust’s Drafts* King’s College, London, 2012. Available online: [http://research.cch.kcl.ac.uk/proust\\_prototype/about.html](http://research.cch.kcl.ac.uk/proust_prototype/about.html).
- 40 Malte Rehbein, Patrick Sahle, and Torsten Schaßan, editors. *Kodikologie und Paläographie im Digitalen Zeitalter - Codicology and Palaeography in the Digital Age*. Schriften des Instituts für Dokumentologie und Editorik. Books on Demand, Norderstedt, 2009. Available online: <http://kups.ub.uni-koeln.de/volltexte/2009/2939/>.
- 41 M. Reisz. Surfdom. *Times Higher Education Supplement*, 2028:34–39, 8 December 2011. Available online: <http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=418343&c=2>.
- 42 *Rights Reserved. - Free Access. Nutzungsrechte für historische Quellen und Handschriften im Internet*. Available online: <http://www.infoclio.ch/de/node/26076>.
- 43 Lambert Schomaker. Advances in writer identification and verification. *Proc. of 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007)*, 26 September, Curitiba, Brazil, 2:769–773, 2007.

- 44 D. Sculley and Bradley M. Pasanek. Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4):409–424, 2008. doi: 10.1093/lc/fqn019.
- 45 Y. Choueka, L. Wolf, R. Shweka, and N. Dershowitz. Automatic extraction of catalog data from digital images of historical manuscripts. Forthcoming.
- 46 Charles Percy Snow. *The Two Cultures*. Cambridge University Press, 2012.
- 47 F. M. Stenton. The supremacy of the Mercian kings. *English Historical Review*, 33(132):433–452, 1918. doi: 10.1093/ehr/XXXIII.CXXXII.433.
- 48 Peter A. Stokes. Palaeography and image processing: Some solutions and problems. *Digital Medievalist*, 3, 2007/8. Available online: <http://www.digitalmedievalist.org/journal/3/stokes/>.
- 49 Peter A. Stokes. Computer-aided palaeography: Present and future. In Rehbein et al. [40], pp. 309–338, 2009. Available online: [http://kups.ub.uni-koeln.de/volltexte/2009/2978/pdf/KPDZ\\_I\\_Stokes.pdf](http://kups.ub.uni-koeln.de/volltexte/2009/2978/pdf/KPDZ_I_Stokes.pdf).
- 50 Peter A. Stokes. Palaeography and the “virtual library” of manuscripts. In B. Nelson and M. Terras, *Digitising Medieval and Early Modern Culture*, pp. 137–169. Medieval and Renaissance Texts and Studies, 426. Arizona Center for Medieval and Renaissance Studies, Tempe AZ, 2012.
- 51 Peter A. Stokes. Modeling medieval handwriting: A new approach to digital palaeography. In *DH2012 Book of Abstracts*, ed. by Jan Christoph Meister et al., pp. 382–385. University of Hamburg, 2012. Available online: <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography>.
- 52 Peter A. Stokes. *English Vernacular Script from Æthelred to Cnut, ca 990 – ca 1035*. Boydell, Woodbridge, Forthcoming.
- 53 D. Stutzman. AAA – AΔΛ – Alphabet, Ambiguité et Actualité (paléographique): l’ontologie des formes alphabétiques. *Paléographie Médiévale*, 2012. Available online: <http://ephepaleographie.wordpress.com/2012/01/22/aaa-%CE%B1%CE%B4%CE%BB-alphabet-ambiguite-actualites-paleographique-ontologie-formes-alphabetiques/>.
- 54 Tagung. *Codicology and Palaeography in the Digital Age*. Conference report, July 2009. Available online: <http://www.i-d-e.de/events-des-ide/2009-tagung-kpdz/report>.
- 55 Simon Tanner. *Measuring the Impact of Digital Resources: The Balanced Value Impact Model*. King’s College, London, 2012. Available online: <http://www.kdcs.kcl.ac.uk/innovation/impact.html>.
- 56 M. Terras. *Digital Images for the Information Professional*. Ashgate, Aldershot, England, 2008.
- 57 *Transcribe Bentham*. Available online: <http://www.transcribe-bentham.da.ulcc.ac.uk/>.
- 58 Georg Vogeler. Kodikologie und Paläographie im digitalen Zeitalter / Codicology and palaeography in the digital age. *AHF-Information*, (206), 2009. Available online: <http://www.ahf-muenchen.de/Tagungsberichte/Berichte/pdf/2009/206-09.pdf>.
- 59 C. Von Oertzen and K. Wilder. *Scholarly Publishing and the Issues of Cultural Heritage, Fair Use, Reproduction Fees and Copyrights*. Berlin, Max Planck Institute, 2008. Available online: <http://www.mpiwg-berlin.mpg.de/PDF/MPIWGWorkshop1-2008Report.pdf>.
- 60 *What’s the Score*. Available online: <https://www.zooniverse.org/lab/wts/>.
- 61 Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1978–1990, 2011.
- 62 European Science Foundation. *Exploratory Workshop on Digital Palaeography*. Conference report, 2011. Available online: <http://hsozkult.geschichte.hu-berlin.de/tagungsberichte/id=3968>.