The Living Web Observatory

Web science encompasses the study of the Web as a network comprising practically limitless quantities of information and nearly unlimited numbers of users, based on a highly decentralised, multifacetted and flexible communication and information infrastructure. Human communication and cooperation are reflected on the Web, alongside many aspects of production, business and economics, and leisure time in our society. Within the scope of the L3S Challenge "The Living Web Observatory", we examine all of the questions that help advance all aspects of the study of data, people, and interactions on the Web.

For our work on this challenge to be successful, we need to solve a number of problems:

**Data collection**

Data collection and archiving. Because the structure of the Web is decentralized and driven by various different interests, regularly collecting and compiling the data needed for interesting studies is a trivial matter. Data gathering and collection is facilitated by the development of Web archive infrastructures, along with cross-institutional initiatives, especially the Web Observatory Initiative launched by the Web Science Trust and W3Tech Labs, and by the W3C’s Web Observatory Community Group. Putting together and providing relevant collections of data and digital artifacts of all kinds is another important aspect of this field, as is development of infrastructures and organizational processes for collecting these kinds of data stocks. Ultimately, infrastructures used to support Web communities and associated activities having to do with collection of use data (as in the case of the BitSonomy system, for example), enable interesting analyses.

**Data preparation**

Preparation and linking of data. The second step is the preparation, including semantic preparation, of data (meaning not only unstructured text and multimedia content, but also semi-structured content, linked data, and conventional structured data along with selection, cross-linking, and contextualization thereof within the scope of relevant communities and interactions, and with an eye to appropriate criteria, in order to enable relevant studies. Developing, documenting, and establishing these processes is an important task for the burgeoning research field of Web science.

**Aggregation and visualization.** Arriving at insights based on the data that have been collected presupposes that there are appropriate mechanisms in place to make it possible to handle and analyse the data. These mechanisms include appropriate dynamic forms of aggregation along various dimensions and suitable forms of visualization. For analysis purposes, the actual analysis and interpretation of those data. The term "Living Web Observatory" thus points to the Web’s importance as a platform for experimental research in the field of Web science.

Finally, an important aspect of the Living Web Observatory is the ability to explore not only the present, but also the past of the Web, as significant parts of our cultural heritage are produced and consumed online. In this context, traditional institutions keeping our cultural heritage need to be complemented with facilities for preservation and public access of online cultural assets, so that we can retrieve, explore and analyse all relevant content available through the Web now, in the past and in the future.

Forgotten and preservation. Since the Web is a core part of nearly every area of our lives, and because it is growing and changing to a significant degree, researchers are increasingly asking what kinds of information can and should be preserved, and what can and should be "forgotten" instead. This field also gives rise to interesting research questions in the area of controlled digital forgetting. Another closely associated factor is the interdisciplinary question of whether there is a right to forgetting on the Web, meaning whether information on people (photos, for instance), should simply disappear from the Web after a certain period in order to limit the negative influence of long past "guilty indiscretions" on current processes such as applying for a new job.

**Data protection law**

PRISM, NSA and European Data Protection Law

For months now the assertion is circulating that there exists a comprehensive monitoring program which allows U.S. intelligence agencies to collect and preserve parts of the Web, the most prominent one being the Internet Archive, which has collected more than 3 Petabytes of Web-content since 1996. Two important national libraries engaged in Web preservation are the British Library and the German National Library, with the aim to preserve national Web content. However, the means for accessing and exploring these archives are severely underdeveloped. None of the archive initiatives is able to provide their collections through an interface which comes close to the functionalities we see today on modern Web search engines.

Relying on Web search engine technology for Web archive retrieval only helps to a certain extent. While providing a more sophisticated interface, the underlying retrieval model of Web search engines still assumes a rather static collection, and does not really take the evolution of the Web into account, with content and link structures changing over time, and new versions of pages added continuously. In addition, the change of entities and their references over time impacts content retrievability.

We need better ways to provide entity-based access to Web archives, taking the evolution of entities and content into account as well as the incompleteness and inconsistency of entity information itself. We also need principled models of temporal entity-based retrieval building on these models, which are also able to connect social media information to conventional Web content. Finally, we need concepts and infrastructures for complex and long-running exploration and analysis processes, complementing the short and non-persistent search sessions most relevant for today’s Web search.

L3S researchers will investigate these issues in depth in our new project ALEXANDRIA, financed through the European Research Council for the next five years with an overall budget of about 2.5 Mill. Euro.

The issue currently occupies, among others, the European Parliament, the European Commission and also inner-German Politics. It is not only due to PRISM very timely. Rather, for months now Europe
What are the goals of web archiving at the German National Library and how can you define its social mandate?

Collecting the German cultural and scientific heritage, indexing and preserving it and making it accessible has always been the mandate of the German National Library. However, when in 2006 the Law on the German National Library extended this mandate to include digital publications, it added a new dimension to our mission.

Since 2006, publications in German and published in Germany, including recorded music and sheet music, is collected, indexed, archived not only in printed format or on other physical carriers, but also in digital format if distributed as purely web-based publications, be it from commercial publishers and distributors or distributed as open web resources.

We will gradually include more subject areas. In addition we collect websites related to predefined events, as for example the last parliamentary elections. Currently a call for tender is out for a complementary Top-Level-Domain-Crawler to collect all websites with the URL “.de” – within given technological limits.

In your opinion, what are the differences in archiving print media in comparison to archiving web publications and web content?

Websites change continuously and, hence, can be only secured as momentary snapshots. Therefore a complete archiving is not realistic. Moreover, due to technical limitations, not all content can be collected. This is especially true for interactive, dynamic formats. Another difference is the sheer amount of web publications that clearly exceeds print publications. The dynamics of the web also means websites appearing and disappearing all the time.

How could the web archive of the German National Library be used in future?

Due to current copyright law we can only allow access to our web archive within our reading rooms. It is however possible to search for websites in the library’s catalogues. Personally, I hope that in the future we will be given the mandate to give access to the web archive under predefined conditions to selective users outside our reading rooms. We are working continually to improve our collections and our retrieval services, qualitatively as well as qualitatively, for example in cooperation with L3S.

Are there any open questions or issues in the context of web archiving the German National Library is currently concerned about?

There are open questions in nearly every respect: collecting the websites, search and retrieval options for the archived websites as well as the long-term availability. There are still many open issues and room for improvement. The German National Library is aware of these issues and is working on solutions - in cooperation with competent partners.

One challenging question for instance concerns the retrieval of older web material. How can websites from the dawn of the Internet be searched using state-of-the-art retrieval tools which are adapted to today’s linguistic usage? It is amazing how fast language is changing and evolving on the internet.

Which cooperation is important for the German National Library in the area of web archiving?

The German National Library cooperates with partners within the International Internet Preservation Consortium (IIPC) to find solutions that can be applied internationally. On the national level we cooperate in an expert’s network for long-term archiving called rektor and in the research group “Documentation and Archiving of Websites” associated with the research group for economic management (WiW). The German National Library also tries to develop its know-how and to improve the quality of its web archiving by working together with service-providers. And, last but not least, by working together with research partners in defined projects because we need to find solutions!

Which problems can you anticipate concerning IPR (Intellectual Property Rights), which solutions are possible?

IPR restricts the re-use of websites even if they are freely accessible on the site of their creator. It is totally unrealistic to expect individual rights clearance for all these masses of open websites. It would certainly help enormously to be provided with a legal solution which would allow making our web archive publicly accessible after a certain period of time and for research purpose. In combination with an option for the creators to block internet access to archived websites, this would mean a considerable balance between the right of the creator and the right of the public to get access to the once openly available records of the web as we know it today. But for now, we have to begin by archiving comprehensively and by creating attractive services – there is still plenty of work to be done!

The German National Library is entrusted with the task of collecting, permanently archiving, bibliographically classifying and making available to the general public all German and German-language publications from 1993, foreign publications about Germany, translations of German works, and the works of German-speaking emigrants published abroad between 1933 and 1945. The German National Library maintains co-operative relations on the national and international level. It is, for instance, the leading partner in developing and maintaining rules and standards in Germany and plays a significant role in the development of international standards.

It is a federal institution with legal capacity under public law. The annual funds provided from the budget of the Federal Government Commissioner for Culture and the Media currently amount to roughly EUR 46 million. The German National Library was preceded by several institutions: these included the Deutsche Bucherei, founded in Leipzig in 1912, and the Deutsche Bibliothek Frankfurt am Main, established in 1946. The German Music Archive (Deutsches Musikarchiv) was founded in Berlin and has been part of the German National Library since 1970, it has been based in Leipzig since December 2010. More information: http://www.dnb.de/EN/wir/
New Web Science Projects

**ERC Advanced Grant ‘Alexandria’**

For his project “ALEXANDRIA: Foundations for Temporal Retrieval, Exploration and Analytics in Time-Related Archives” Professor Wolfgang Nejdl, Head of the L3S Research Center, has received an ERC Advanced Grant, starting in 2014. ERC Advanced Grants allow exceptional established researchers to pursue ground-breaking, high-risk projects that open new directions in their respective research fields or domains. The European Research Council will fund ALEXANDRIA for five years, with an overall budget of about 2.5 Mill. Euro.

The project is named after the Library of Alexandria, in Egypt, which was one of the largest and most significant libraries of the ancient world. The following paragraphs give a short summary of the ALEXANDRIA project.

Significant parts of our cultural heritage are produced on the Web, yet only insufficient opportunities exist for accessing and exploring this vast body of the Web. The ALEXANDRIA project aims to develop models, tools and techniques necessary to archive and index relevant parts of the Web, and to retrieve and explore this information in a meaningful way. While the easy accessibility to the current Web is a good baseline, optimal access to Web archives requires new models and algorithms for retrieval, exploration, and analytics which go far beyond what is needed to access the current state of the Web. This includes taking into account the unique temporal dimension of Web archives, structured semantic information already available on the Web, as well as social media and network information.

**ERC Grant ‘Individualized Implant Placement – IIP’**

Prof. Dr. Bodo Rosenstiel received an ERC-Proof-of-Concept Grant for the Project “IIP”. The project is an on-top funding for the ERC-starting Grant “Dynamic MiniIIP”. The goal of the project is a software development and start-up which allows to individually place implants for a virtual surgery planning procedure. Based on implant placement recommendations, misplacements, pain and reduced mobility after surgery will be reduced in the future. The developed methods are based on a statistical database analysis, which will be linked to several projects at the L3S.

According to the Bundesverband Medizintechnik (German Medical Technology association), Germany is the third biggest market for medical technologies in the world. The primary motivation for the project is to help overcome the current fragmented approach taken by health systems in European countries, which makes gaining access to a consent record of individual citizens that involves cross-border activities difficult.

The MHA vision is to generate a 4D avatar representing the health status of patients and in the process to encourage the engagement of both medical professionals and patients. The avatar will be designed as a lifetime companion for individual citizens that facilitates the collection of, and access to, long-term health status information. By representing health status in this way, it is hoped to make a valuable contribution to individualized disease prediction and prevention, moreover, greater awareness of healthy lifestyles and the achievement of independent living will be promoted.

The MHHealthAvatar (MHA) project is a research and demonstration-oriented iterative venture, in which the feasibility of creating personal health records configured as digital patients will be studied. The primary motivation for the project is to help overcome the current fragmented approach taken by health systems in European countries, which makes gaining access to a consent record of individual citizens that involves cross-border activities difficult.

The MHHealthAvatar will be built on the latest ICT technology with an aim of engaging public interest to facilitate the collection of, and access to, long-term health status in this way, it is hoped to make a valuable contribution to individualized disease prediction and prevention, moreover, greater awareness of healthy lifestyles and the achievement of independent living will be promoted.

**FORGETIT – Forgetting and Preserving: Friends or Foes?**

Unprecedented amounts of digital content are created today in professional and personal life, triggered e.g. by digital photography and social networks. This raises the question on how to best deal with these masses of material on the long run transforming it from raw stored material into assets, which remain usable, understandable and also enjoyable even over decades. Just sufficient storage does not solve the problem: formats and technologies evolve over time, hard disks break, life situations change, etc. leading to both random and increasing losses over time.

Inspired by the processes of forgetting and remembering in the human brain, the European project ForgetIT (coordinated by L3S), is creating technology for intelligent preservation of personal and organizational content. For this purpose, a Managed Forgetting process is introduced, which translates the power of human forgetting for staying focused into digital memory. This includes a variety of forgetting options such as summarizing and reducing redundancy, and helps to decide about investments into preservation actions. Furthermore, Synergistic Preservation technology will enable a smooth transition from active information use into preservation, thus lowering the barrier for preservation technology use. Finally, technology for Contextualized Remembering keeps archived information usable and understandable, even if accessed after decades.

**BRENDA – Comprehensive Enzyme Information System**

The BRENDA enzyme represents the world wide largest information system of functional biochemical and molecular enzyme data with more than 3 million manually annotated data (from more than 125,000 references) on enzyme function, occurrence, kinetics, and molecular properties. 190,000 small molecules or other ligands (substrates, cofactors, inhibitors etc.) interacting with the enzymes can be searched and displayed.

BRENDA provides access to seven interconnected databases, and each entry is linked to a reference, the source organism, and if available the protein sequence. The BRENDA Tissue Ontology is an encyclopaedia of organs, tissues, and cell types. The content is compiled by further databases (AMENDA, FRENDA, BRENDA, KENDA) which add literature information on enzyme occurrence, enzyme-related diseases and kinetic values, based on elaborated text-mining methods. More than 3000 users visit the BRENDA website each day, mainly from North America, Asia, and Europe. BRENDA is freely available to the scientific community.

BRENDA is essential for a large community of researchers in the molecular life sciences ranging from basic research in molecular biology and systems biology to medical applications, drug development, and biotechnology.

**WEB GOVERNANCE – WG**

**E-SCIENCE – ESC**

**INTERACTIVE ACCESS TO INFORMATION – IA**

**NEXT GENERATION INTERNET – NGI**

**E-SCIENCE – ESC**
In August 2013, L3S Research Center put its new computing cluster into operation. The cluster allows researchers to store and analyze datasets of up to 60 terabytes size – an amount of data that corresponds to 22 billion pages of paper, which – if stacked on top of each other – would result in a tower of 2200 kilometers in height. Such datasets can accumulate easily in Web Science, e.g., just by crawling the web pages of German universities, researchers at L3S already collected 6 gigabytes of data. Since the hardware setup that is available in typical computing centers is not able to process such large amounts of data efficiently, L3S decided to acquire its own cluster. The system consists of nine machines with 96 processor cores and a distributed disk space of almost 200 terabytes. Using special software – based on the MapReduce paradigm originally developed at Google – L3S researchers are now leveraging the cluster to analyze large datasets, e.g., more than one billion tweets from Twitter. The cluster is of special importance for the ALEXANDRIA project in which L3S will develop novel approaches for the exploration and analysis of web archives. An extension is already planned for 2014 – it will considerably increase both the storage capacity and the networking bandwidth.

The recent vision of the “Future Internet” attracts significant networking research and causes controversial debates on the actions to be taken. Clean-slate initiatives envision a fresh start that puts fundamental principles of networking into question. Avoiding any constraints of the current Internet implementation, the ambition of the clean-slate approach is to understand and design the “right” network architecture. Evolutionary approaches, on the other hand, seek incremental improvements, assuming that the Internet can – as in the past – be fixed to accommodate the changing needs of users and applications.

Dagstuhl hosted a three-day seminar on the Future Internet on March 25-27, 2013 organized by Jon Crowcroft (University of Cambridge), Klara Nahrstedt (member advisory board L3S, University of Illinois - Urbana), Ralf Steinmetz (member advisory board L3S, TU Darmstadt), and Markus Fidler (member L3S, Leibniz Universität Hannover). At the seminar, around 40 invited researchers from academia and industry discussed the promises, approaches, and open challenges of the Future Internet. The seminar focused on topics in the areas of prescriptive network theory, experimentally-driven research, software defined networking, virtualization, and OperaFlow. During the seminar, the following questions were discussed and (partly) answered: Is a prescriptive network theory feasible? Which insights can the experimental, testbed-based approach reveal? What are the challenges for wide-area service-oriented virtual networks?

L3S Acquires Big Data Computing Cluster

Dagstuhl Seminar on the Future Internet

L3S MEMBERS
Prof. Dr. techn. Wolfgang Nejdl
L3S Executive Director
(Leibniz Universität Hannover)

Prof. Dr. Wolf-Tilo Balle
L3S Director (TU-Braunschweig)

Prof. Dr. Nikolaus Forró
L3S Director (Leibniz Universität Hannover)

Prof. Dr. Ing. Markus Fidler
(Leibniz Universität Hannover)

Prof. Dr. Andreas Holto
(Julius Maximilian Universität Würzburg)

Prof. Dr. Robert Jaschke
(Leibniz Universität Hannover)

Prof. Dr. Ing. Marcus Magnor
(TU Braunschweig)

Prof. Dr. Panagiotis Papadimitriou
(Leibniz Universität Hannover)

Direktor Uwe Rosemann
(ITB Hannover)

Prof. Dr. Bodo Rosenhahn
(Leibniz Universität Hannover)

Prof. Dr. Dietmar Schomburg
(TU-Braunschweig)

Prof. Dr. rer. nat. Matthew Smith
(Leibniz Universität Hannover)

Prof. Dr. Irina Sens (ESC)

Prof. Dr. Stefan Dietze (IAI)

Prof. Dr. Thomas Risse (AI4)

Prof. Dr. Robert Jaschke (AI4)

Prof. Dr. Karl Aberer
(WeST) - Universität Koblenz, Koblenz, Germany

Prof. Dr. Christoph Meinel
München, Germany

Prof. Dr. Dieter Kranzlmüller
University of Illinois, USA - Urbana, IL

Prof. Dr. Karl Aberer
Science (I&C), Lausanne, Switzerland

Prof. Dr. Ing. Gabriele von Voigt
(Leibniz Universität Hannover)

DAGM-Prize 2002, Dr.-Ing. Siegfried Werth Prize

His works received several awards, including a DAGM-Prize 2002, Dr.-Ing. Siegfried Werth Prize 2003, DAGM-Main Prize 2005, IVCNZ best student paper award , DAGM-Main Prize 2007 and Olym- pus-Prize 2007. In 2011, the European Commission awarded Bodo Rosenhahn with a 1.43 million Euros ERC-Sting Grant. He published more than 100 research papers, journal articles and book chapters. His current citations are ERC-Starting Grant. He published more than 100 research papers, journal articles and book chapters. His current citations are in the area of model development and over 60 book-volumes are in the area of modeling and experimental characterization of biocatalysis and metabolism. His group is developing the enzyme information system BRENDA since 1987. Prof. Schomburg was a member of the L3S in 2013.

L3S Members

Bodo Rosenhahn

Prof. Bodo Rosenhahn studied Computer Science (minor subject Medicine) at the University of Kiel. He received the Dipl.-Inf. and Dr.-Ing. from the University of Kiel in 1999 and 2002, respectively. From 10/2003 till 10/2005, he worked as PostDoc at the University of Auckland (New Zealand), funded with a scholarship from the German Research Foundation (DFG). In 11/2005-08/2008 he worked as senior researcher at the Max-Planck Institute for Computer Science in Saarbrücken. Since 09/2008 he is Full Professor at the Leibniz-Universität of Hannover, heading a group on automated image interpretation.

His works received several awards, including a DAGM-Prize 2002, Dr.-Ing. Siegfried Werth Prize 2003, DAGM-Main Prize 2005, IVCNZ best student paper award , DAGM-Main Prize 2007 and Olympus-Prize 2007. In 2011, the European Commission awarded Bodo Rosenhahn with a 1.43 million Euros ERC-Sting Grant. He published more than 100 research papers, journal articles and book chapters. His current citations are ERC-Starting Grant. He published more than 100 research papers, journal articles and book chapters. His current citations are...