OKKAM – Enabling the Web of Entities

Entities such as persons, organizations, locations, and products are prevalent in many of today’s information sources. They are described and commented on, for example, in databases, Web pages, and news articles, and Blogs. As a consequence, a variety of information about individual entities exists. However, the ad hoc integration and combination of such information is often hindered by naming variants, ambiguities, and spelling mistakes. Therefore, it is the goal of the OKKAM project to establish an infrastructure for re-using unique entity identifiers or names, i.e., an Entity Name System. The widespread use of avatar-like entity names eases information integration and enables a variety of useful entity-centric services, such as entity-centric search.

A “universal resolver” which provides DNS (domain name system) functionality on the level of entities has been identified by Bradley Horowitz in a BBC article [1] as a necessity for enabling the “Internet of Things” or Web of Entities. In the Web of entities – as opposed to the current Web – useful entities are the unit of consideration. This is in contrast to the commonplace document-centric viewpoint, in which the document – possibly containing useful information about an entity – is the unit of search.

The Entity Name System (ENS), which is developed in OKKAM, is such a universal resolver for entity names. Building upon a large entity repository and advanced entity matching methods, it enables the user to find “the” unique identifier for an entity – the OKKAM ID, if the entity has already been described by somebody else. The use of a matching framework ensures the flexibility and effectiveness of the infrastructure. It can be extended with new matching modules, which are, for example, customized to domain and entity-type specific characteristics. By integrating OKKAM IDs into the respective description of entities in an information source, the respective source is prepared for supporting ad hoc integration with other sources, and for enabling entity-centric services. It is no longer necessary to perform complex analysis for that two sources speak about the same entity; with OKKAM IDs, it is obvious from the use of the same entity name. The OKKAM ENS is accompanied by a set of useful tools and services: The OKKAM-empowered tools, for example, support the creation of content that already includes OKKAM IDs for the referenced entities. The most prominent example is a Word Plugin developed by the Italian information extraction company, Experts Systems which – on the fly-recognizes entities and suggests OKKAM IDs for annotation of the entities. This obviously employs information extraction technology in addition to the ENS functionality.

Prominent companies such as SAP and the publisher Elsevier have joined OKKAM, because they understand the added value that entity-centric services can provide for many of their customers. Entity-centric applications developed in OKKAM include an advanced editor for scientific authors – with a strong involvement of Elsevier, which recognizes and handles domain specific entities (e.g., proteins and genes). SAP – within OKKAM – focuses on the development of entity and product-centric solutions for organizational knowledge management and customer support.

OKKAM – focuses on the development of entity extraction technology in addition to the ENS functionality

Knowledge, Information, Learning @ L3S

The L3S Research Center focuses on innovative methods and technologies for three key research fields of the European Information Society, namely Knowledge, Information and Learning. L3S plays a very active role in European research, and is one of the key players in the areas of digital libraries and intelligent content, internet technologies and search, as well as technology enhanced learning. Third party funding through the EU/IST program on digital libraries is at one third of our nearly 5.5 Million Euro yearly budget. A similar amount comes from German funding agencies and from industry.

Digital Libraries

Digital libraries has been a topic where L3S members have been active since L3S’s inception. Recent projects focus on innovative solutions for elicience (GoDI-Grid) and on new services for libraries and better access to digital information. These include LINSearch, which enables advanced catalog and abstract access for technical information and ViaChem, which is improving search and the use of digital information in the area of chemistry. Finally, the LWa project, which started this year and is funded within the EU/IST program on digital libraries, is an example for advanced digital preservation and digital library aspects in the context of Web archives.

Social Media

Social Media has emerged as a dominant phenomenon on the World Wide Web. When compared to edited content by commercial and other classical information providers, user-generated content is representing an ever increasing share of the Web’s information. L3S is very active in that area, with social media and social network research playing a prominent role in the NPEMUKE and PAKATES projects. L3S members have also been active in contributing to European and national strategic documents on these topics, including a recent European report by the Networked Media Unit of DG Information Society and Media of the European Commission in the future of Web 2.0.

Future Internet

A new relevant area is Future Internet, which is mentioned as one of the important federating research themes in the FP7 IST work program 2007–2013. This theme involves digital content within all projects relevant for Digital Libraries and other digital collections, but also the increasing connection and integration of real and virtual worlds (“Internet of Things”) where L3S has valuable expertise through its projects on robotics and mobile communications.
The European Library/ Europeana Net

Jill Cousins is Director of the European Library and Programme Director of Europeana.eu. The digital library, museum and archive for Europe that is being developed by the thematic network known as EDLnet. She began at the national library of the Netherlands, the Koninklijke Bibliotheek in 2004, and brought the European Library service into operation. Due to the success of this service, the European Union has given their strategic backing to The European Library for the creation of Europeana.eu.

LiWA – Living Web Archives

Due to the central role that the World Wide Web plays in nearly all areas of today’s life, adequate Web archiving has become a cultural necessity in preserving knowledge. First generation Web archiving technology has been built by pioneers in the domain based on existing search technology. The next generation Web archiving technologies will overcome limitations in content capture, preservation, analysis and enrichment. The aim of LiWA is to reach authentic, trusted sources and to see them juxtaposed with each other. This gives serendipity a cross border, serious chance. It also means, for example, that the user finding a reference to a piece of jewellery that was bought for Hermitia Mary Sauer’s 21st birthday by William II – and seen in a letter from Constantijn Huygens to Caspar Du Bois in Antwerp – will also be able to see the van Dyck picture of the marriage of William and Mary, with Mary wearing the piece of jewellery. Or, a user can find the original score of a piece of music in a French archive, but will also play the piece through a German music museum. In a way it is less about fulfilling their expectations; users don’t miss what they never had. Rather, the goal is to try to make sure that they get real material, not just a second hand unauthenticated version. The site also brings to the web some things that have never seen the light of day and allows a virtual repatriation of material.

LiWA: What is the difference between national libraries and the European Library?

Hm...not quite sure what this question is getting at, but national libraries are showing collections and their information in the context of their own countries and national remits. The European Library aggregates available content for users to do a single search and retrieve information held in all the national libraries.

LiWA: Who are the users of EDL – the European digital library?

What is the benefit for them in using EDL and what are their expectations?

EDL is now called Europeana. It is distinct from The European Library, as it addresses all the cultural content holders across four domains of museums, libraries, archives and audio visual collections. It is owned by the EDL Foundation, which is a legal entity representing the major associations involved in cultural heritage across Europe such as EURICA, ICCU, CESAC, IFLA, FIAT, ILASS, LIBER, CENL.

We have no users at the moment, since the site is in the process of being created, but the aim is to attract the general user, who is looking for fun, for tourists, for interest as well as individuals who want to conduct academic research. For the prototype in November 2008, we will only have one interface, but for the operational service, we are considering maintaining different interfaces for different user groups and allowing particular groups to create their own versions of Europeana. For instance, the European Blind Union would take the data and its underlying structures and create a site for the visually impaired. Alternatively, a school education site would mash up relevant bits of Europeana with their own resources: such tasks will be realized through web services or APIs.

I think the benefit is the ability to reach authenticated, trusted sources and to see them juxtaposed with each other. This gives serendipity a cross border, serious chance. It also means, for example, that the user finding a reference to a piece of jewellery that was bought for Hermitia Mary Sauer’s 21st birthday by William II – and seen in a letter from Constantijn Huygens to Caspar Du Bois in Antwerp – will also be able to see the van Dyck picture of the marriage of William and Mary, with Mary wearing the piece of jewellery. Or, a user can find the original score of a piece of music in a French archive, but will also play the piece through a German music museum. In a way it is less about fulfilling their expectations; users don’t miss what they never had. Rather, the goal is to try to make sure that they get real material, not just a second hand unauthenticated version. The site also brings to the web some things that have never seen the light of day and allows a virtual repatriation of material.

LiWA: What are the challenges to building a European digital library?

These are human, political and technical challenges. The technical one is quite serious, given the scale of the project. Many of the nice semantic things would not be realised, as yet, have no proof of concept that they are scalable for the volumes we are considering. We envision making the collections and data from the four domains – museums, archives, libraries and audio visual collections – interoperable; however, it will take more manual mapping than I would like to see for future maintenance. On the human side, the challenge is in getting buy-in from the different stakeholders, with different views for the use and purpose of their resources. A museum, for example, likes to tell the story, so the curatorial aspects are important. On the other hand, an archivist wants to ensure that the content of the document is understood, while libraries just want to provide access. This necessitates a long process of consensus building. Finally, from a political point of view, we have real support from the Commission, but this has to be transferred to the countries if we are to have long term sustainability.

LiWA: What are Web Search engines the portal to world wide information space. What is the relationship of Europeana to Web search?

Europeana will make use of the web search engines as channels for users to access its content. It will actively promote content to the web search engines so that users can find the site and its interconnections.

LiWA: The idea of Web 2.0 is changing from passive web consumers to active contributors. How will this influence the library world?

The Web 2.0 can influence the library world in a very positive manner. For example, it can fill in the gaps, and probably more than they are really ready for at the moment; since it means re-lasishing control and trusting the user to do what is right, and not thinking that users will destroy or corrupt work. In theory, it could be a symbiotic relationship whereby users are helping the libraries to give greater, broader access and also to filling in the gaps in the explanations, or the metadata. Users themselves are also good at providing the links. This would be particularly useful in Europeana, for example, in making the link between the archive letter and the object of art.

LiWA: Libraries are worried about loss of control and corruption of their intellectual property. I think we can separate the user generated content from that of the professionally contributing institutions and still end up with a much stronger and more accessible resource.

LiWA: Recently a number of publication servers for online journals became popular. Do you make this content accessible?

Do you mean those owned by the publishers themselves? If so, then yes, we would aim to give access to this content, though probably not in the prototype, as initially we are not dealing with rights or payment, but we do have a project proposal called EuropeanaConnect which brings in OUP from the outset.

LiWA: Let’s talk about different types of content. Is Europeana providing pictures, videos and audio documents?

It accesses them from the provider site via links. It pulls together the data to be able to find information in an index and allows the user to search across multiple media. It will not hold content itself, but leave it on the servers of the providers – thereby allowing all types of content to be represented, ultimately even including 1D and 3D data.

LiWA: Some national libraries are also archiving their countries’ web domain. Will this content also be included in Europeana?

No, I don’t know yet as this is subject to all sorts of personal data law restrictions, so the national libraries may archive it but not give access and this varies from country to country. We will have to wait and see how this evolves although the goal would be to access it.

LiWA: Can you give us a vision for the future? How will Europeana Library look in 20 years?

This is very difficult. I think in three years time it will probably still be a stand alone site for the national libraries. However, it is like an environment that is it offers a wider geographical area than Europeana. Much depends on whether European can be made sustainable in the long term and the implementation of a workable business model. If this happens, then maybe The European Library will not only have a web presence, but also a dark portal continuing to aggregate content for the national libraries and serving Europeana or other library related sites. Equally, this could be completely accessible and The European Library will be thriving in twenty years time.

LiWA: Jill Cousins thank you very much for this interview.
Personalized Information Spaces for Chemical Libraries

Today’s challenge in information retrieval is no longer limited to the find-and-search process for the right resources in a digital library, but to that of finding the right piece of information within a plethora of returned search results. Integrated search-ports are able to simultaneously search in multiple and heterogeneous document collections containing millions of documents. Information providers like digital libraries are challenged by the information overload resulting in the availability of more and more information. VfAChem II aims to assist chemists in their search for the right piece of information by creating metadata enriched documents. We present an integrated search and navigation through query results taking into account individual retrieval strategies, thus forming a personalised knowledge space of chemistry.

Nowadays a scientist can easily find an abundance of information for many areas of research, facing the challenge of finding the important information within literature, as something essential. Consider, for instance, a pharmaceutical researcher starting a research project involving a G-protein-coupled receptor (GPCR). Before 1988 there would only have been a few journal articles to read based on a search across the essential databases for chemical and related sciences. By 1992 there was already a total of 290 relevant documents to read. However, by 2005 the researcher would be confronted with 14000 articles dealing with GPCRs.

The use of data mining methods on large document collections offers new ways to assist and guide a user on his way to the right piece of information. In chemistry and its related areas information is typically about molecules - their properties, analytics, synthesis, reactions, biological and pharmaceutical activity, industrial use and more. Generally information about chemical entities is communicated in two major ways: either as language-specific text or as graphical chemical structure information. Documents in chemistry, thus may contain chemical information in text and images in multiple variations. The VfAChem II project explores mining chemical document collections by using methods for chemical entity recognition, recognition of reaction types or optical chemical structure recognition. The challenge of chemical named entity recognition can be demonstrated by a commonly used substance named acetylsalicylic acid. Although there is a general accepted standard for naming compounds from the Union of Pure and Applied Chemistry (IUPAC) one may find other names, such as acetylsalicylate or 2-Acetylsalicylic acid, all accepted and understood by a chemist. However, most readers will know the described substance by its trade name Aspirin. Once a chemical entity is recognised, it can be converted into the corresponding chemical structure information based on the rules of systematic chemical names. Moreover, all extracted chemical entities can be used in the generation of ontologies of chemical entities or reactions, making use of the hierarchical representation of substance classes in chemistry.

VfAChem II will also offer different access paths to information. Besides simple keyword search it makes use of the faceted browsing paradigm. A semantic taxonomy generation algorithm which dynamically creates personalized facets. Using the relationship between documents and salient keywords this arranges relevant keywords in a hierarchical fashion. The information retrieval interface of VfAChem II hence will provide the user with a new experience in searching large document collections. Combined with Web 2.0 features like social tagging, the use of ontologies, taxonomies, and facets enable the user to navigate through query results based on his personalized knowledge space.

BibSonomy – A Social Bookmark and Publication Sharing System

Are you also bothered by the need to report your publications several times, in different formats, to different institutions? And do you use yet another system – or no system at all – for managing the publications that you read and cite? Would you like to share more information about these publications with your colleagues and co-workers? Then you might want to consider the social bookmarking system BibSonomy. This Web 2.0 service is freely accessible at http://www.bibsonomy.org/

BibSonomy is a service for managing and sharing web pages and publications on the Internet. There, you can centrally store the references to your publications, as well as add keywords to those resources. The keywords, so-called tags, permit you to structure your own collection of publications as well as to search for entities of other users: additionally, you can store bookmarks for web pages. Using BibSonomy, one can search the literature and bookmarks of other participants with similar interests and discover new, interesting web sites. This social perspective allows personalized recommendations, a feature that global search engines, such as Google, cannot provide, since they are not aware of their users’ interests. By offering interested communities (e.g. university groups or groups of the German society for computer science) the creation of user groups in BibSonomy to enable teams to organize their internal and external literature exchange.

The data structure of BibSonomy – a so-called folksonomy – offers the possibility to add more than one tag to a resource, so that documents are related to multiple topics. This is in contrast to ‘authority control systems’ in which each documentation entry is assigned to a single authority. Folksonomy provides a new way to explore resources with a semantic search engine. BibSonomy is often compared to ‘semantic web’, where documents are linked to other published content by their metadata (tags). This system is valuable especially for starting researchers, who want to share their work, and for the public, which is interested in discovery-led navigation.

The ViFaChem II project explores mining chemical document collections by using methods for chemical entity recognition, recognition of reaction types or optical chemical structure recognition. BibSonomy is a social bookmarking system which support the search and analysis of data in document collections. Combined with Web 2.0 features like social tagging, the use of ontologies, taxonomies, and facets enable the user to navigate through query results based on his personalized knowledge space.

Research Perspectives

Apart from offering a modern service to manage and exchange knowledge, a major motivation for the development of BibSonomy was its applicability as a research platform. We developed methods which support the search and analysis of data in social bookmarking systems. These approaches are based on the tripartite hypergraph of a folksonomy, which is created when users add keywords to their contents. For instance, we adapted association rules to the folksonomy structure which can be deployed to learn tag hierarchies for a better structuring of tags. Our FolkRank permits to rank tags, users and resources. The algorithm depends on user preferences so that related elements of a folksonomy appear on the top of a ranking. The approach can also be applied for detecting communities of interest in folksonomy. Our publications on BibSonomy and related topics can be downloaded from http://www.bibsonomy.org/group/kele-myuwen.

German National Library of Science and Technology (TIB) Hannover

As the German National Library of Science and Technology, TIB specialises in all areas of technology and their related sciences such as architecture, chemistry, information technology, mathematics and physics. Founded in 1955, TIB has been successfully fulfilling its role as an important source of information for decades. With an excellent stock of 6 million items, as well as 18,000 current journals, the TIB today stands as the largest specialised library in the world, and one of the most reliable suppliers of documents within its field of specialisation.

The collection's strengths are conference reports, research papers, patent applications, standards, PhD theses as well as specialist literature from
Distributed User Modeling in the GRAPPLE Project

The GRAPPLE project aims at delivering a technology-enhanced learning environment which guides users through a lifelong learning experience. A key underpinning of the GRAPPLE system design is its automatic adaptation to a user’s personal preferences, prior knowledge, skills competences, learning goals, and personal or social context.

Information technology is widely used in learning processes. Most institutions of higher education and many larger corporations use some kind of learning management systems (LMS) for processes related to the delivery of learning material. However, deploying a one-size-fits-all LMS does not automatically lead to better learning processes. The adaptation of technology-enhanced learning processes to the needs of individual users is a key factor to the widespread use and success of learning technology within organizations.

The main objective of GRAPPLE is to support lifelong learning by means of a personalized and adaptive, technology-enhanced learning environment that is seamlessly integrated with major learning management systems in a service-oriented web-based approach.

Furthermore, it is our goal that GRAPPLE will create a responsive learning environment that not only motivates, engages and inspires learners, but also effects better learning results. Research in the area of personalization will be used as a key component in achieving the desired goal.

The L3S Research Center is mainly involved in the work on distributed user modeling methods and techniques. In cooperation with DKFZ, the Technical University of Eindhoven, and Delft University of Technology we are designing a platform and storage architecture that is open for both internal and external user-adaptive systems. Rather than adopting a monolithic approach, as was common in the 1990s, the core system will provide basic functionality for connected parties to share and reason about each others’ refined and contextualized user profiles via standardized interfaces. In theory, this could mean that one will have to choose between several competing parties that offer their own versions of a ‘user interest model’. A provenance mechanism, which enables tracing back all reasoning steps back to the original observation, allows parties to verify the trustworthiness and applicability of the inferences. We expect that this competitive, evolving, and demand-driven approach will address the needs of adaptive learning environments far better than a predefined, centralized user model could do.

The GRAPPLE consortium largely consists of members of the Adaptive Hypermedia and User Modeling communities. During the Eindhoven kick-off meeting, the shared background and well-established personal relationships played an indispensable role in jumpstarting the effort. The next GRAPPLE consortium meeting will be held in Hannover, Germany, prior to the Adaptive Hypermedia 2008 conference.

L3S Members

Prof. Dr. techn. Wolfgang Nejdl (Leibniz University Hannover) Executive Director

Prof. Dr. med. Gabriele von Voigt (Leibniz University Hannover) Director

Prof. Dr. Wolf-Tilo Balke (TU Braunschweig) Director

Prof. Dr.-Ing. Klaus Jobmann (Leibniz University Hannover) Director

Prof. Dr. Gerald Stumme (University of Kassel) Director

Prof. Dr.-Ing. Bernardo Wagner (Leibniz University Hannover)

L3S Associate Members

Prof. Dr.-Ing. Christian Grimm (Leibniz University Hannover)

Prof. Dr. Nicola Henze (Leibniz University Hannover)

Prof. Dr.-Ing. Thomas Kaiser (Leibniz University Hannover)

Scientific Advisory Board

Prof. Marc Eisenstadt Knowledge Media Institute, The Open University, United Kingdom

Prof. Dr. Matthias Jarke Fraunhofer Institute for Applied Information Technology, Bonn Department of Computer Science V, Aachen University of Technology, Germany

Prof. Dr. H. c. mult. Hermann Maurer Institute for Information Systems and Computer Media, Graz University of Technology, Austria

Prof. Dr. Erich J. Neuhold Faculty of Computer Science, University of Vienna, Austria

Prof. Dr. Thomas Ottmann Institute for Computer Science, Albert-Ludwigs-University of Freiburg, Germany

Prof. Dr. H. c. mult. A.-W. Scheer Institute for Information Systems, University of Saarbrücken, Germany

Prof. Gis Wiederhold Computer Science Department, Stanford University, USA

Legal Information:

Dr. Uwe Thaden
(Managing Director)
L3S Research Center
Appelstr. 9a
30167 Hannover, Germany
Tel.: +49 (0) 511 762 - 1771
Fax: +49 (0) 511 762 - 1779
E-Mail: thaden@L3S.de
Internet: www.L3S.de

Prof. Dr. Wolfgang Nejdl is elected as first Editor-in-Chief for IEEE TLT

Prof. Dr. Wolf-Tilo Balke has accepted a full professorship for databases and information systems at the University of Braunschweig together with the directorate of the Institute for Information Systems (http://www.ifis.cs-bbs.de/). Prof. Balke has been member of the L3S for three years contributing mainly in the research on databases and digital libraries and now strengthens the bond between the L3S Research Center and the University of Braunschweig.

L3S Research Center at the CeBIT 2008

The L3S Research Center was present at CeBIT 2008, “future part”, Hall 9 Stand B22 as part of a joint exhibition of Lower Saxony research institutions. This year’s feature was the NEPOMUK project aiming at the Social Semantic Desktop.

NEPOMUK’s main objective is to transform your desktop into an intelligently interlinked workplace: in daily office work, a plethora of documents and emails are sent and received. Documents are created and revised in preparation for meetings and reports. These objects are stored in different locations, in various directories of a user’s PC or in email folder structures. As such, an overview to present, demonstrate, and discuss the latest ideas about the Web, its infrastructure, relevant standards bodies working to shape the Web.

URL: http://www.grapple-project.org
Contact: Prof. Dr. Nicola Henze, henze@L3S.de
Dr. Eike Herder, herder@L3S.de

Best student award for Sukriti Ramesh

Sukriti Ramesh, a visiting student from Vellore Institute of Technology India (VIT) received the best student award from her home university. Sukriti is currently on a six-months research visit at L3S, writing her Bachelor thesis on query optimization for distributed databases. The best student award is the highest student distinction given at VIT University and it is accompanied by a gold medal and a scholarship.

Prof. Dr. Wolfgang Nejdl will chair the 18th International World Wide Web Conference (http://2009.org) together with Yvonne Maarek from Google. The WWW Conference is a global event that brings together key researchers, innovators, decision-makers, technologists, businesses, and standards bodies working to shape the Web. Organized since 1994, the conference series is the premier venue for academics and industry to present, demonstrate, and discuss the latest ideas about the Web, its infrastructure, relevant algorithms and new innovative applications.

It is most of all an annual opportunity for the international community to discuss and debate the state and the evolution of the Web.

The technical program for the five-day conference, which will take place in Madrid in April 2009, will include refereed paper presentations, plenary sessions, panels, and poster sessions, in the areas of User Interfaces and Mobile Web, Data Mining, Industrial Practice and Experience, Internet Monetization, Performance and Scalability, Rich Media, Search, Security and Privacy, Semantic/Data Web, Social Networks and Web 2.0, Web Engineering as well as XML and Web Data. The WWW2009 program will also include Tutorials and Workshops, a “WWW in Iberoamerica” track, a W3C track, a Developers track, and Exhibitions.