Proceedings of the European Data Forum 2012
June 6-7, 2012, Copenhagen DK

Collected Abstracts of Posters and Demonstrations

Edited by
Michael Hausenblas, DERI, Ireland
Elena Simperl, Karlsruhe Institute of Technology, Germany

Contents

1 Poster: "Semantic Search: Technologies and Case Studies"

Christoph Goller, INTRAFIND Software AG

Abstract. The amount of data that is available in digital form is growing exponentially. Analyzing large data sets (so-called big data) will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus. A considerable amount of data is produced as unstructured text. Therefore, search, search-based applications, and text analytics will play an important role in the new emerging data economy. I will briefly explain the most important text analytics methods from Morphological Analysis and POS-Tagging to Information Extraction, Named Entity Recognition and Text Classification and show how these methods can be used to generate information/semantics from unstructured text automatically. Furthermore I will present case studies from commercial customer projects showing how text analytics can enable semantic applications such as automatic newsletter generation, product classification, semantic linking, expert identification, semantic search and even question answering.

2 Demonstration: “Graphity - Generic Linked Data Platform”

Martynas Jusevicius, graphity.org

Abstract: Graphity is a platform for publishing and analysis of open Linked Data as well as for consumption and integration of heterogeneous datasources as Linked Data. Enabled by radically extensible software platform (built on W3C standards and open-source code) and customizable user-friendly interfaces, Graphity allows seemless mashups between RDF datasources such as DBPedia and Web 2.0 services like Twitter and Basecamp. We aim to connect popular Web 2.0 services to the ever-growing LOD cloud and to provide integrated platform for editing, visualizations, analysis, reports, and import/export of Linked Data. This will be a step towards the unified semantic "data locker", lowering barriers of entry for SMEs and engaging both citizen end-users and a

community of contributors and developers. We expect the Graphity platform to have a positive impact on transparency, emerging societal and technological trends such as data science, data journalism, infographics, crowd analytics, and data-mining.

3 Poster: "Industry Perspectives on EUDAT - A Pan-European Common Data Infrastructure"

David Manset1 and the EUDAT Consortium, MAAT France

[1]MAAT France, Argonay, France; dmanset@maatg.fr
[2] http://www.eudat.eu; eudat-pmo@postit.csc.fi

Abstract. A relentless trend of massive data generation is happening across all human activities. It is estimated that the amount of data produced each year is greater than the sum of all that previously created, thus exceeding ICT tools and capacities. Although the challenges may seem daunting and ambitious to address, the opportunities are immense. With a proper infrastructure and accompanying set of services in place, researchers, engineers and users may be able to share data and exploit it to its full potential, to derive new knowledge and develop innovative applications.

Within this landscape, the EUDAT project aims to contribute to the production of a pan-European Collaborative Data Infrastructure (CDI) addressing data proliferation in Europe's scientific and research communities. Its vision is to support a CDI, which will allow researchers to share data within and between communities and enable them to carry out their daily work effectively. EUDAT aims to provide a solution that will be affordable, trustworthy, robust, persistent and easy to use, over time. It offers the opportunity for communities to contribute to and benefit from a distributed, managed, sustainable and production quality service hosting framework.

EUDAT is thus focusing on the development of six foundational services, which are considered as top priorities for the communities and which address the core functional requirements of the CDI. The proposed presentation will therefore introduce the EUDAT CDI services and elaborate on the industrial per-

spectives, which the latter open in associated communities and industry sectors.
Keywords. E-infrastructure, Big Data, CDI, Commodity Service, Digital Object Identifier, Cloud, Grid, HPC

4 Demonstration: "Connected Media Experiences"

Lyndon Nixon, STI International Consultancy und Research GmbH

Abstract: The demo will show how structured data published on the Web can be used to interlink media resources and generate automated enrichments of video.

5 Poster: "Image analysis technologies on large dataEnvironments"

Artzai Picon[1], Arantza Bereciartua1, Sergio Rodriguez1, Angel Lopez1,Elena Muñoz[2], Fabienne Gandon[3], Francesco Moscone[4], Peter H.J. Riegman[5], Sonsoles García[6], Roberto Bilbao [7]

[1]FUNDACIÓN TECNALIA RESEARCH & INNOVATION, Parque Tecnológico de Bizkaia, Edificio 202, 48170, Zamudio, Bizkaia, Spain,
[2]EMEDICA S.L., Ribera de Axpe 11 D1, 48950, Erandio, Bizkaia, Spain,
[3]PERTIMM (PERTINENT ET IMMEDIAT) SAS, 51, Boulevard Voltaire, 92600, Asnières-Sur-Seine, France
[4]BRUNEL UNIVERSITY, Kingston Lane, Uxbridge, Middlesex, UB8 3PH, United Kingdom
[5]ERASMUS UNIVERSITAIR MEDISCH CENTRUM ROTTERDAM, Gravendijkwal 230, 3015 CE, Rotterdam, The Netherlands
[6]CULTEK S.L.U., Av. Cardenal Herrera Oria, 63, 28034, Madrid, Spain
[7]FUNDACIÓN VASCA DE INNOVACIÓN E INVESTIGACIÓN SANITARIAS (BIOEF), Plaza Asua s/n , Sondika, Bizkaia, Spain

Corresponding author: artzai.picon@tecnalia.com

Abstract: Visual information has been assumed to be the most important source of information in human beings. This assertion can be transferred to the digital universe where different types of visual information sources coexist. Pictures on social networks, images of virtual encyclopedias, corporate sites, museums, parks, contain images that allow us to contextualize and extract

information from their content. This embedded information is related to who our friends are and activities we like, as well as evoking feelings. However, these images are composed of sets of pixels that show a great variability that precludes the use of traditional methods for data analysis, especially in applications involving large volumes of data. The proper modeling of the illumination effects, the color models, the extraction of specific visual features or morphologies, or even the use of information in the non-visible spectrum of color allows us to extract and analyze this visual information appropriately.

In this poster we present the techniques used to allow a correct description of images for different applications such as image retrieval, object detection and characterisation. This allows us to develop different types of applications such as contextualization of digital content, similar image search (image retrieval) in many types of fields arising from histological image characterisation to the detection of near-duplicates in copyrighted content through the detection of specific objects in images.

6 Poster: "Semantic Search: Technologies and Case Studies"

Artazai Picon Picon[1], Arantza Bereciartua[1] , Elena Muñoz[2], Fabienne Gandon[3], Francesco Moscone[4], Peter H.J. Riegman[5], Sonsoles García[6], Roberto Bilbao[7], INTRAFIND Software AG

Services associated to digitalised contents of tissues in biobanks across Europe – BIOPOOL

[1]FUNDACIÓN TECNALIA RESEARCH & INNOVATION, Parque Tecnológico de Bizkaia, Edificio 202, 48179, Zamudio, Bizkaia, Spain,
[2]EMEDICA S.L., Ribera de Axpe 11 D1, 48950, Erandio, Bizkaia, Spain,
[3]PERTIMM (PERTINENT ET IMMEDIAT) SAS, 51, Boulevard Voltaire, 92600, Asnières-Sur-Seine, France
[4]BRUNEL UNIVERSITY, Kingston Lane, Uxbridge (Middlesex), UB8 3PH, United Kingdom
[5]ERASMUS UNIVERSITAIR MEDISCH CENTRUM ROTTERDAM, 's Gravendijkwal 230, 3015 CE, Rotterdam, The Netherlands
[6]CULTEK S.L.U., Av. Cardenal Herrera Oria, 63, 28034, Madrid, Spain
[7]FUNDACIÓN VASCA DE INNOVACIÓN E INVESTIGACIÓN SANITARIAS (BIOEF), Basque Biobank for Research-O+Ehun Plaza Asua s/n , Sondika, Bizkaia, Spain

Abstract: Nowadays it has become common practice to take digital images of thin slices of biopsies that are obtained for studying the composition of cells, glands, tissues and organs, and the possible pathologies that may affect them. These images are of high interest in medical diagnostics, research and education.

Pathology departments in hospitals and Biobanks are usually the facilities that provide archived biological samples for use in life sciences. They manage the tissue samples, the associated digital images and other complementary digital data (health information such as pathologies or treatments followed by the patient). Although most of the pathology departments and biobanks adequately capture and store digital images of the different biologic materials, it is not so usual for biobanks to adequately associate the representative health information to the digital images of their samples and it is even less usual sharing these images in a network. So, the digital images are usually spread all over different systems stored in different formats, databases and facilities belonging to different types of institutions and they are not easily identifiable and reachable. This creates a difficult environment for sharing and reusing this type of data between different interested organisations.

The BIOPOOL consortium project has been created to carry out a new approach, that arises from the need of pathology departments and biobanks of sharing, exchanging, processing, understanding and exploiting the digital histology images and the data associated to the biologic material stored in these institutions. The project will develop the needed technology to extract and gather this digital information from different pools, analyse it, and being able to compare it and to score images similar to one provided as a search pattern based on an innovative Content Based Image Retrieval (CBIR) system capable of searching histological images using different mixed text and image queries. BIOPOOL will establish a complete intelligent biobank and pathology department network, building a constructive basis for pan-European cooperation in diagnosis and medical research.

Seven partners from four different countries collaborate together in BIOPOOL. A great effort is done by the SMEs within this consortium project, The results coming from BIOPOOL will be translated into new services and products using

the technologies and data pools provided by the other organisations allowing more precise actuations in digital pathology, especially in diagnosis.

7 Poster: "Getting on the Data Highway"

Luis Rodrigues, Tenforce

Abstract: TenForce is a Belgian software company. Established in 2001 it has been active as a software service company in the areas of internet, publishing, semantic technology, meta data management, Linked Open Data. TenForce will reveal some of the threats and opportunities of the dataweb through specific examples and projects it has been working on. Attendees will learn what this new paradigm means for different stakeholders: existing publishers, government, service developers.
The session will also zoom in on the components and mechanism of a typical dataweb or linked open data solution, already available today. Attendees will understand the basis architecture for dataweb or linked open data solutions. It will become clear that metadata, availability of metadata on the web and linking of metadata are cornerstones of any solution. TenForce will present some case studies from real live projects. Through these case studies you will learn the typical approach and methodology to deploy data web and linked open data solutions. You will learn how to mitigate risks and guarantee success.

8 Poster: "Real-time data analytics of drilling sensor streams for prediction of critical situations"

Herwig Zeiner, Bernhard Jandl, Martin Winter, Roland Unterberger, Rudolf Fruhwirth, Christian Derler

Abstract: The aim of this real-time sensor data analytics system is to make better decisions and to predict and avoid upcoming critical situations in drilling operations. The purpose is to support the drilling engineers with the ongoing processes and give them deeper insights into the current situations. This

should help to avoid upcoming critical situation. It is of utmost importance to prevent damages to equipment, protect the crew from injuries, and avoid environmental pollution in drilling operations nowadays. We present a real-time data analytics prototype. The proposed research prototype has been applied to real world data of oil or gas reservoirs in onshore regions as well as in offshore regions. The research prototype consists of a complete data processing chain including several modules for data acquisition from sensors on the drilling platform, adaptive sensor data analytics and problem specific visualization. While the data acquisition modules collect the data from sensors at the rig and produce a live data stream in an appropriate WITS like format, the data processing algorithms have to analyze the data streams in real-time and classify the drilling operations, detect potentially abnormal upcoming critical events and give appropriate advice to the drilling crew, if possible. Finally, all the raw sensor data streams as well as several adaptive online learning algorithms results and several sensor channel quality parameters of the rig are visualized in a novel user interface to support drilling employees at the rig. Following that the current drilling situation is presented in a comprehensive manner and in real-time.