Number 111 October 2017

ERCIM NEWS

www.ercim.eu

Special theme:

Digital Humanities

Also in this issue:

Research and Innovation:

Trend Analysis of Underground Marketplaces

Editorial Information

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 6,000 printed copies and is also available online.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
Tel: +33 4 9238 5010, E-mail: contact@ercim.eu
Director: Philipp Hoschka, ISSN 0926-4981

Contributions
Contributions should be submitted to the local editor of your country

Advertising
For current advertising rates and conditions, see http://ercim-news.ercim.eu/ or contact peter.kunz@ercim.eu

ERCIM News online edition
http://ercim-news.ercim.eu/

Next issue
October 2017, Special theme: Digital Humanities

Subscription
Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu or by filling out the form at the ERCIM News website: http://ercim-news.ercim.eu/

Contents

JOINT ERCIM ACTIONS

SPECIAL THEME

The special theme section "Digital Humanities" has been coordinated by George Bruseker (ICS-FORTH), László Kovács (MTA-SZTAKI), and Franco Niccolucci (University of Florence)

Digital Source Indexing and Analysis

Information Management Strategy

Research Infrastructure Development

3D Analysis Techniques

Information Visualization and Communication

RESEARCH AND INNOVATION

This section features news about research activities and innovative developments from European research institutes

EVENTS, IN BRIEF

Call for Proposals

Report

In Brief

ERCIM Membership

After having successfully grown to become one of the most recognized ICT Societies in Europe, ERCIM has opened membership to multiple member institutes per country. By joining ERCIM, your research institution or university can directly participate in ERCIM's activities and contribute to the ERCIM members' common objectives playing a leading role in Information and Communication Technology in Europe:

- Building a Europe-wide, open network of centres of excellence in ICT and Applied Mathematics;
- Excelling in research and acting as a bridge for ICT applications;
- Being internationally recognised both as a major representative organisation in its field and as a portal giving access to all relevant ICT research groups in Europe;
- Liaising with other international organisations in its field;
- Promoting cooperation in research, technology transfer, innovation and training.

About ERCIM

ERCIM – the European Research Consortium for Informatics and Mathematics – aims to foster collaborative work within the European research community and to increase cooperation with European industry. Founded in 1989, ERCIM currently includes 15 leading research establishments from 14 European countries. ERCIM is able to undertake consultancy, development and educational projects on any subject related to its field of activity.

ERCIM members are centres of excellence across Europe. ERCIM is internationally recognized as a major representative organization in its field. ERCIM provides access to all major Information Communication Technology research groups in Europe and has established an extensive program in the fields of science, strategy, human capital and outreach. ERCIM publishes ERCIM News, a quarterly high quality magazine and delivers annually the Cor Baayen Award to outstanding young researchers in computer science or applied mathematics. ERCIM also hosts the European branch of the World Wide Web Consortium (W3C).

> **“Through a long history of successful research collaborations in projects and working groups and a highly-selective mobility programme, ERCIM has managed to become the premier network of ICT research institutions in Europe. ERCIM has a consistent presence in EU funded research programmes conducting and promoting high-end research with European and global impact. It has a strong position in advising at the research policy level and contributes significantly to the shaping of EC framework programmes. ERCIM provides a unique pool of research resources within Europe fostering both the career development of young researchers and the synergies among established groups. Membership is a privilege.”**
>
> ***Dimitris Plexousakis, ICS-FORTH, ERCIM AISBL Board***

Benefits of Membership

As members of ERCIM AISBL, institutions benefit from:

- International recognition as a leading centre for ICT R&D, as member of the ERCIM European-wide network of centres of excellence;
- More influence on European and national government R&D strategy in ICT. ERCIM members team up to speak with a common voice and produce strategic reports to shape the European research agenda;
- Privileged access to standardisation bodies, such as the W3C which is hosted by ERCIM, and to other bodies with which ERCIM has also established strategic cooperation. These include ETSI, the European Mathematical Society and Informatics Europe;
- Invitations to join projects of strategic importance;
- Establishing personal contacts with executives of leading European research institutes during the bi-annual ERCIM meetings;
- Invitations to join committees and boards developing ICT strategy nationally and internationally;
- Excellent networking possibilities with more than 10,000 research colleagues across Europe. ERCIM's mobility activities, such as the fellowship programme, leverage scientific cooperation and excellence;
- Professional development of staff including international recognition;
- Publicity through the ERCIM website and ERCIM News, the widely read quarterly magazine.

How to Become a Member

- Prospective members must be outstanding research institutions (including universities) within their country;
- Applicants should address a request to the ERCIM Office. The application should inlcude:
  - Name and address of the institution;
  - Short description of the institution's activities;
  - Staff (full time equivalent) relevant to ERCIM's fields of activity;
  - Number of European projects in which the institution is currently involved;
  - Name of the representative and a deputy.
- Membership applications will be reviewed by an internal board and may include an on-site visit;
- The decision on admission of new members is made by the General Assembly of the Association, in accordance with the procedure defined in the Bylaws (http://kwz.me/U7), and notified in writing by the Secretary to the applicant;
- Admission becomes effective upon payment of the appropriate membership fee in each year of membership;
- Membership is renewable as long as the criteria for excellence in research and an active participation in the ERCIM community, cooperating for excellence, are met.

Please contact the ERCIM Office: contact@ercim.eu

ERCIM "Alain Bensoussan" Fellowship Programme

ERCIM offers fellowships for PhD holders from all over the world. Topics cover most disciplines in Computer Science, Information Technology, and Applied Mathematics. Fellowships are of 12 months duration, spent in one ERCIM member institute. Fellowships are proposed according to the needs of the member institutes and the available funding.

Application deadlines for the next round: 30 April and 30 September 2017

More information: http://fellowship.ercim.eu/

HORIZON 2020 Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 80 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach.

The ERCIM Office has recognized expertise in a full range of services, including identification of funding opportunities, recruitment of project partners, proposal writing and project negotiation, contractual and consortium management, communications and systems support, organization of attractive events, from team meetings to large-scale workshops and conferences, support for the dissemination of results.

How does it work in practice?
Contact the ERCIM Office to present your project idea and a panel of experts will review your idea and provide recommendations. If the ERCIM Office expresses its interest to participate, it will assist the project consortium as described above, either as project coordinator or project partner.

Please contact:
Philippe Rohou, ERCIM Project Group Manager
philippe.rohou@ercim.eu

HRADIO – A New Project to Tackle Future of Internet Radio

ERCIM/W3C is participating in HRADIO (Hybrid Radio everywhere for everyone), a new EU-funded project (Innovation Action) that started in September 2017. It focuses on radio service innovations enabled by convergence. While radio, with its rich editorial content, remains a highly popular medium, listening figures are slowly declining, particularly among youngsters. With the rapid rise of smartphones, radio faces competition from many new services including music steaming platforms. Regular radio today often does not include attractive features as known from online platforms. And if present, they are mostly not well integrated with the actual radio programme. This is where HRADIO will deliver.

Driven by the industry need to create attractive new radio experiences, the project will leverage the potential of hybrid technology for radio, enabling the integration of cost-effective broadcast distribution with the web. Broadcasters will be enabled to personalise radio services (while respecting privacy), to provide intuitive functionalities like time-shifting and, eventually, to foster and to exploit user engagement. HRADIO will pave the way to bring these features not only to broadcasters' native mobile applications, but also to portals, to connected radios and into the car. The core approach is to integrate validated solutions and to harmonise APIs which together will provide broadcasters with an abstracted service layer accessible across any device and distribution platform – ensuring sustainability and return of investment. Therefore, consumers will be able to access their personal radio services on different devices and platforms enabled by a seamless broadcast-internet integration for radio content distribution. Eventually, HRADIO will publish its developments as ready-to-use Android and HTML client implementations.

W3C will lead the standardisation activities of the project, working with partners to align modelling and main technical tasks with relevant standards, and to identify and specify possible extensions to standards from which the project would benefit. W3C will thus contribute to the definition of scenarios and requirements, the overall system architecture and specifications. W3C will also lead the management of innovation, dissemination, communication and exploitation plan work package of the project.

The project has a duration of 30 months from 1 September 2017 to 29 February 2020 with an overall budget of 3,25 Mio Euro. Project partners include VRT, Belgium; Institut für Rundfunktechnik GmbH, Germany; Ludwig-Maximilians-Universität München, Germany; GEIE ERCIM France; RBB, Germany; Konsole Labs GmbH, Germany; UK Radioplayer Ltd.

Link:
http://cordis.europa.eu/project/rcn/211079_en.html

Please contact:
Francois Daoust, W3C, fd@w3.org

Introduction to the Special Theme

by George Bruseker (ICS-FORTH), László Kovács (MTA-SZTAKI), Franco Niccolucci (University of Florence)

This special theme "Digital Humanities" of ERCIM News is dedicated to new projects and trends in the interdisciplinary domain between computer science and the humanities.

Just as experimental research is based on reasoning over experiments, research in the humanities is based on reasoning over sources, which may be textual, material or intangible. Digital humanities (DH) is the interdisciplinary research field that studies how computer science may support such investigations by creating tools and tailoring computer technology to the specific needs of humanities research while also addressing the methodological issues that arise in relation to the adoption of an intensive use of such digital means. This issue is dedicated to showcasing cutting edge work being undertaken in the domain of DH in the areas of: digital source indexing and analysis, information management strategy, research infrastructure development, 3D analysis techniques, and information visualisation and communication.

Digital Source Indexing and Analysis

Following on from an earlier period of research and investment in digitisation, when huge amounts of analogue content were turned into digital products, the research frontier in DH has now shifted to an interest in automatic or semi-automatic ways of dealing with digital sources. Such sources include both the aforementioned digitised materials and also, increasingly, born-digital texts and other digital media such as audio, images and video. A key research challenge is to create methodologies and tools for finding the needle in the proverbial haystack of millions of poorly indexed files. At present, hot topics include the Optical Character Recognition (OCR) of digitised manuscripts and the parallel techniques of Speech Recognition, as well as the use of Named Entity Recognition (NER) on the resulting digital text files. This is the subject of the article, "In Codice Ratio: Scalable Transcription of Vatican Registers", by Firmani et al., which proposes a supervised NER using mixed algorithmic and crowdsourcing approaches, as well as of Felicetti's paper, "Teaching Archaeology to Machines: Extracting Semantic Knowledge from Free Text Excavation Reports" and Brouwer's, "MTAS – Extending Solr into a Scalable Search Solution and Analysis Tool on Multi-Tier Annotated Text". Two articles address multimedia sources, presenting innovative solutions to retrieval and discovery. Dologlou at al. deal with other audio and visual sources in their, "Phonetic Search in Audio and Video Recordings" while Köhler et al. address speech recognition and analysis in the contribution, "KA3: Speech Analytics for Oral History and the Language Sciences". Addressing the general question of querying and discovery in large datasets in the humanities, Devezas et al. introduce a perspective and strategy on use of integration of information using graph technology, "Graph-Based Entity-Oriented Search: Imitating the Human Process of Seeking and Cross Referencing Information". Another increasingly important topic of research arising in this area relates to fact checking and determining the veracity of claims made in sources as well as impact of arguments in social contexts. The contributions of Manolescu in, "ContentCheck: Content Management Techniques and Tools for Fact-checking" and Heder et al. in, "Argumentation Analysis of Engineering Research" offer perspectives on how to critically assess structured resources through pattern recognition.

Information Management Strategy

The questions of the long-term control of this 'needle in the haystack' problem and the efficient use of research data sources raise general methodological issues of how to create robust information management strategies in the humanities. This area of research addresses the rapid expansion of digital sources in multiple formats that cover both broader and more precise topics and the question how we can create long-term sustainable access for the reuse of resources in a way that promotes accessibility and the quality necessary to support academic research. Part of the question here is how to create and successfully use common or transparent expressions/translations of data across domains, as well as how to create, share and properly use common vocabularies for describing data. Advances in the areas of vocabularies and semantics are reported in Daskalaki's et al., "A Back Bone Thesaurus for Digital Humanities" and in Bruseker et al., "Meeting the Challenges to Reap the Benefits of Semantic Data in Digital Humanities". Beyond the data question, however, lie practical and social dimensions to the problem of long-term data management, a question of understanding and formalising procedures and protocols in a new digital research environment and putting them efficiently in place in information systems in the present research economy. Clivaz et al. in, "HumaReC: Continuous Data Publishing in the Humanities" and Basset's, "A Data Management Plan for Digital Humanities: the PARTHENOS Model" provide views and answers on how to meet the challenges, obligations and opportunities that arise for digital humanists creating digital archives that are re-usable by others in an open access framework. Another area of important research is in supporting trust in digital resources. This area is taken up by van Ossenbruggen in, "Trusting Computation in Digital Humanities Research". Finally, a perennial area demanding new development and imagination lies in creating tools that allow researchers to generate data meeting the above criteria in a non-onerous manner.

Research Infrastructure Development

In order to support forward looking, comprehensive and critical approaches to these issues, the European Commission supports a great part of the funding for digital humanities research, especially through the Infrastructures Programme. In particular, this programme supports the creation of research infrastructures (RI), which are networks of facilities, resources and services offered to a research community to support and catalyse their work. Following the roadmap created by the ESFRI (European Strategy Forum on Research Infrastructures), some of these RIs are upgraded and designated as an ERIC (European Research Infrastructure Consortium), according to the recommendation of a panel of experts. An ERIC is a transnational institution tasked with managing an RI and fostering innovation in the related research field. This issue brings news on the progress of a number of ERICS within the humanities domain. Edmond's et al. report on the activity of DARIAH (Digital Research Infrastructure for the Arts and the Humanities) in, "The DARIAH ERIC: Redefining Research Infrastructure for the Arts and Humanities in the Digital Age", while the digital infrastructure of E-RIHS (European Research Infrastructure for Heritage Sciences) is described in Pezzati's, "DIGILAB, a New Infrastructure for Heritage Science". Meanwhile, Bardi et al. present research into the design and implementation of a generalised information architecture for digital humanities in, "Building a Federation of Digital Humanities Infrastructures". Finally, in "Knowledge Complexity and the Digital Humanities: Introducing the KPLEX Project", Edmond et al. announce interesting new research in the context of such RIs taking an explicitly humanities grounded critical look on the concept of 'data' in the first place and how this affects what may or may not be digitized, and the manner in which it is perceived or used.

3D Analysis Techniques

Turning to the application of digital methods to the study of material culture (e.g., man-made objects or architecture), we find continuous innovation in the application of digital techniques in order to try to better understand extant objects or to produce academically sourced and grounded representations of now lost heritage. In this domain, the use of 3D imaging techniques and inventing means of analytically applying these to heritage research is a key area of innovation. In this context, Hanif's et al.'s paper, "Ancient Document Restoration Using Sparse Image Representation", presents a means to link research on documents with research on the matter on which they are recorded, allowing the virtual restoration of ancient documents. Mature research on the application of 3D technology to monuments is represented in Wall's et al., "The Virtual St Paul's Cathedral Project", which reports on the virtual reconstruction of St Paul's based on historical sources. Meanwhile, Barreau's et al., "Immersive Point Cloud Manipulation for Cultural Heritage Documentation" presents an innovative application to use 3D models in a VR platform in order to allow archaeologists to work directly with such products in their research and reporting activities. Finally, in, "Physical Digital Access inside Archaeological Material" Nicolas et al. demonstrate the uses of 3D imaging in conducting non-destructive research on heritage materials. Alliez et al. in, "Culture 3D Cloud" meanwhile present a platform for cloud computing of 3D objects both for enabling their online analysis as well as communicating them to other researchers and the public.

Information Visualization and Communication

Indeed, digital humanities research does not take place in a bubble but, thanks in no small part to its form, is inherently suited to communication to a broader public, both of researchers in other disciplines and generally interested parties. Consequently, considerable research effort is being put into the question of how to present and make digital heritage and humanities content understandable and accessible to this wider audience. Papers addressing the communication of history, heritage or museum exhibits with digital tools in this issue include: Morillo's et al., "Re-Interpreting European History through Technology: The CrossCult Project", and Micsik's et al., "Cultural Opposition in former European Socialist Countries: Building the COURAGE Registry", which look at means of gathering and presenting heretofore under-represented historical information for researchers and the public. A series of contributions also look at innovative ways to explore digital humanities datasets This research takes many directions, from exploring the users of augmented reality as seen in Tamisier's et al., "Locale, an Environment-Aware Storytelling Framework Relying on Augmented Reality", to the application of virtual reality to connect times and space as presented by Koebel's et al., "The 'Biennale 4D' Project" to explorations of new techniques of exploring graph based data on the web as reported by Abrate's et al. in, "The Clavius Correspondence: From Digitization to Visual Exploration of Knowledge". In a related vein, Chessa et al. explore how to connect personal mobile devices to relevant services for data exploration inter alia in, "Service-oriented Mobile Social Networking". Each of these contributions explores innovative approaches to better communicate history and heritage to visitors to museums, monuments or heritage sites around the globe.

The contributions to this special theme issue on digital humanities give an insight into some of the main currents of research currently being undertaken in DH. Such research is carried out as much within the context of smaller research projects as in European funded transnational structures such as the ERICs. The diversity of research demonstrated is strong evidence of the vitality of this interdisciplinary domain, where state-of-the-art digital technology goes hand in hand with the study of human culture of the present and of the past.

Please contact:
George Bruseker
ICS-FORTH, Greece
bruseker@ics.forth.gr

László Kovács
MTA-SZTAKI, Hungary
laszlo.kovacs@sztaki.hu

Franco Niccolucci
University of Florence, Italy
franco.niccolucci@gmail.com

In Codice Ratio: Scalable Transcription of Vatican Registers

by Donatella Firmani, Paolo Merialdo (Roma Tre University) and Marco Maiorino (Vatican Secret Archives)

In Codice Ratio is an end-to-end workflow for the automatic transcription of the Vatican Registers, a corpus of more than 18.000 pages contained as part of the Vatican Secret Archives. The workflow has a character recognition phase, featuring a deep convolutional neural network, and a proper transcription phase using language statistics. Results produced so far have high quality and require limited human effort.

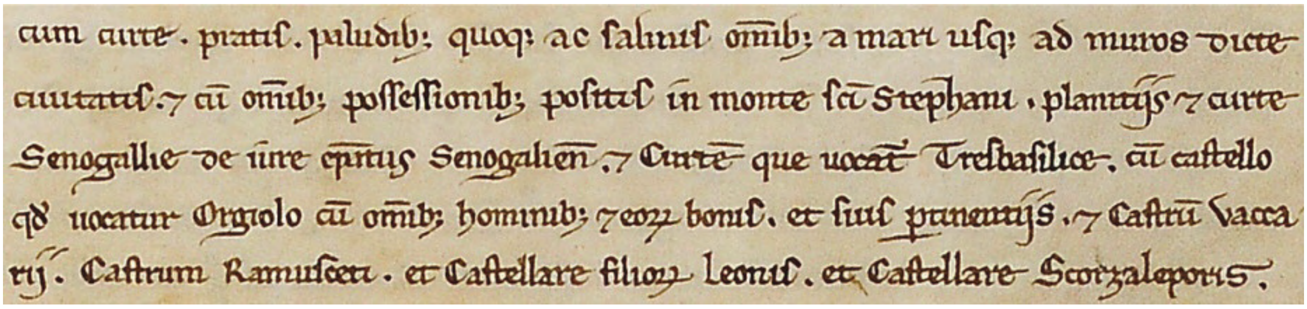
Historical handwritten documents are an essential source of knowledge concerning past cultures and societies [3]. Many libraries and archives have recently begun digitizing their assets, including the Biblioteque Nationale de France, the Virtual Manuscript Library of Switzerland, and the Vatican Apostolic Library. Due to the sheer size of the collections and the many challenges involved in a fully automatic handwriting transcription (such as irregularities in writing, ligatures and abbreviations, and so forth), many researchers in the last years have focused on solving easier problems, most notably keyword spotting. However, as more and more libraries worldwide digitize their collections, greater effort is being put into the creation of full-fledged transcription systems.

Our contribution is a scalable end-to-end transcription workflow based on fine-grained segmentation of text elements into characters and symbols. We first partition sentences and words into text segments. Most segments contain actual characters, but there are also segments with spurious ink strokes. (Perfect segmentation cannot be achieved without transcription. This result is known as Sayer's Paradox.) Then, we submit all the segments to a deep convolutional neural network (CNN), designed following recent progresses in deep learning [2] and the de facto standards for complex optical character recognition (OCR) problems. The labels returned for each segment by the deep CNN are very accurate when the segments contain actual characters, but can be wrong otherwise. Finally, we reassemble such noisy labels into words and sentences using language statistics, similarly to [1].

In Codice Ratio is an interdisciplinary project involving the Humanities and Engineering departments from Roma Tre University, and the Vatican Secret Archives, one of the largest historical libraries in the world. The project started in 2016 and aims at the complete transcription of the "Vatican Registers" corpus. The corpus, which is part of the Vatican Secret Archives, consists of more than 18.000 pages of official correspondence of the Roman Curia in the 13th century, including letters, opinions on legal questions, addressed from and to kings and sovereigns, as well as to many political and religious institutions throughout Europe. Never having been transcribed in the past, these documents are of unprecedented historical relevance. A small illustration of the Vatican Registers is shown in Figure 1.

State-of-the-art transcription algorithms generally work by a segmentation-free approach, where it is not necessary to individually segment each character. While this removes one of the hardest steps in the process, it is necessary to have full-text transcriptions for the training corpus, in turn requiring expensive labelling procedures undertaken by paleographers with expertise on the period under consideration. Our character-level classification has instead much smaller training cost, and allows the collection of a large corpus of annotated data using a cheap crowdsourcing procedure. Specifically, we implemented a custom crowdsourcing platform, and employed more than a hundred high-school students to manually label the dataset. To overcome the complexity of reading ancient fonts, we provided the students with positive and negative examples of each symbol. After a data augmentation process, the result is an inexpensive, high-quality dataset of 23.000 characters, which we plan to make publicly available online. Our deep CNN trained on this dataset achieves an overall accuracy of 96%, which is one of the highest results reported in the literature so far.

The project takes place in Rome (Italy) and features interdisciplinary collaborators throughout Europe and the world, including the Trinity College in Dublin (Ireland), the Max-Planck-Institute for European Legal History in Frankfurt (Germany), and the Notre Dame University in South Bend (Indiana).



cum curte. pratiſ. paludib; quoq; ac ſaltuſ omīb; a mari uſq; ad muros dicte
ciuitatiſ. 7 cū omīb; poſſeſſionib; poſitiſ in monte ſcī Stephani. plantijs 7 curte
Senogallie de iure epātus Senogalien. 7 Curte que uocat Treſbaſilice. cū caſtello
qd uocatur Orgiolo cū omīb; hominib; 7 eoꝝ boniſ. et ſuiſ ptinentijs. 7 Caſtrū Vacca
rij. Caſtrum Ramuſceti. et Caſtellare filioꝝ Leoniſ. et Caſtellare Scorzaleporis.

Figure 1: Sample text from the manuscript "Liber septimus regestorum domini Honorii pope III", in the Vatican Registers.

Domestic collaborators include Sapienza University in Rome (Italy) and two roman high schools, namely Liceo Keplero and Liceo Montale.

The transcriptions produced so far account for lower-case letters and a subset of the abbreviation symbols. Future activities include transcription of abbreviations and upper-case letters, as well as an extensive experimental evaluation of the whole pipeline.

References:

[1] D. Keysers, et al.: "Multi-language online handwriting recognition", IEEE TPAMI, 2017.

[2] D. Kingma, B. Jimmy: "Adam: A method for stochastic optimization", arXiv, 2014.

[3] J. Michel, et al.: "Quantitative analysis of culture using millions of digitized books", Science, 2011.

Please contact:
Donatella Firmani, Paolo Merialdo
Roma Tre University, Italy
+39 06 5733 3229
donatella.firmani@uniroma3.it,
+39 06 5733.3218
paolo.merialdo@uniroma3.it

Marco Maiorino
Vatican Secret Archives
Vatican City State
marco.maiorino2109@gmail.com

Teaching Archaeology to Machines: Extracting Semantic Knowledge from Free Text Excavation Reports

by Achille Felicetti (Università degli Studi di Firenze)

Natural language processing and machine learning technologies have acquired considerable importance, especially in disciplines where the main information is contained in free text documents rather than in relational databases. TEXTCROWD is an advanced cloud based tool developed within the framework of EOSCpilot project for processing textual archaeological reports. The tool has been boosted and made capable of browsing big online knowledge repositories, educating itself on demand and used for producing semantic metadata ready to be integrated with information coming from different domains, to establish an advanced machine learning scenario.

In recent years, great interest has arisen around the new technologies related to natural language processing and machine learning, which have suddenly acquired considerable importance, especially in disciplines such as archaeology, where the main information is contained in free text documents rather than in relational databases or other structured datasets [1]. Recently, the Venice Time Machine project [L1], aimed at (re)writing the history of the city by means of "stories automatically extracted from ancient manuscripts", has greatly contributed to feeding this interest and rendering it topical.

Reading the documents: machines and the "gift of tongues"

The dream of having machines capable of reading a text and understanding its innermost meaning is as old as computer science itself. Its realisation would mean acquiring information formerly inaccessible or hard to access, and this would tremendously benefit the knowledge bases of many disciplines. Today there are powerful tools capable of reading a text and deciphering its linguistic structure. They can work with most major world languages to perform complex operations, including language detection, parts of speech (POS) recognition and tagging, and grasping grammatical, syntactic and logical relationships. To a certain extent, they can also speculate, in very general terms, on the meaning of each single word (e.g., whether a noun refers to a person, a place or an event). Named entities resolution (NER) and disambiguation techniques are the ultimate borderline of NLP, beyond which we can start to glimpse the actual possibility of making the machines aware of the meaning of a text.

Searching for meanings: asking the rest of the world

However, at present no tool is capable of carrying out NER operations on a text without adequate (and intense!) training. This is because tools have no previous knowledge of the context in which they operate unless a human instructs them. Usually, each NLP tool can be trained for a specific domain by feeding it hundreds of specific terms and concepts from annotated documents, thesauri, vocabularies and other linguistic resources available for that domain. Once this training is over, the tool is fully operational, but it remains practically unusable in other contexts unless a new training pipeline is applied. Fortunately, this situation is about to change thanks to the web, big data and the cloud, which offer a rich base of resources in order to overcome this limitation: services like BabelNet [L2] and OpenNER [L3] are large online aggregators of entities that provide immense quantities of named entities ready to be processed by linguistic tools in a cloud context. NLP tools, once deployed as cloud services, will be able to tune into a widely distributed and easily accessible resource network, thus reducing the training needs and making the mechanism flexible and performant.

TEXTCROWD and the European Open Science Cloud

The above is what TEXTCROWD [L4] seeks to achieve. TEXTCROWD is one of the pilot tools developed under EOSCpilot [L5], an initiative aimed "to develop a number of demonstrators working as high-profile pilots that integrate services and infrastructures to show interoperability and its benefits in a number of scientific domains, including archaeology". The pilot developed by VAST-LAB/PIN (University of Florence, Prato) under the coordination of Franco Niccolucci and in collaboration with the University of South Wales, is intended to build a cloud service capable of reading excavation reports, recognising relevant archaeological entities and linking them to each other on linguistic bases [2]. TEXTCROWD was initially trained on a set of vocabularies and a corpus of archaeological excavation reports (Figure 1). Subsequently, thanks to the cloud technology on which it is built, it has been made capable of browsing all the major online NER archives, which means it can also discover entities “out of the corpus” (i.e., non-archaeological ones) and educate itself on demand so that it can be employed in different domains in an advanced machine learning scenario (Figure 2). Another important feature concerns the output produced by the tool: TEXTCROWD is actually able to generate metadata encoding the knowledge extracted from the documents into a language understandable by a machine, an actual “translation” from a (natural) language to another (artificial) one. The syntax and semantics of the latter are provided by one of the main ontologies developed for the cultural heritage domain: CIDOC CRM [L6], an international standard that is very popular in digital humanities.

“Tell me what you have understood”

CIDOC CRM provides classes and relationships to build discourses in a formal language by means of an elegant syntax and to tell stories about the real world and its elements. In TEXTCROWD, CIDOC CRM has been used to “transcribe”, in a format readable by a machine, narratives derived from texts, narrating, for instance: the finding of an object at a given place or during a given excavation; the description of archaeological artefacts and monuments, and the reconstructive hypotheses elaborated by archaeologists after data analysis [3]. CIDOC CRM supports the encoding of such narratives in standard RDF format, allowing, at the end of the process, the production of machine-consumable stories ready for integration with other semantic knowledge bases, such as the archaeological data cloud built by ARIADNE [L7]. The dream is about to become reality, with machines reading textual documents, extracting content and making it understandable to other machines, in preparation for the digital libraries of the future.

Links:

[L1] timemachineproject.eu
[L2] babelnet.org/
[L3] www.opener-project.eu/
[L4] eoscpilot.eu/science-demos/textcrowd
[L5] eoscpilot.eu
[L6] http://cidoc-crm.org
[L7] http://ariadne-infrastructure.eu

References:

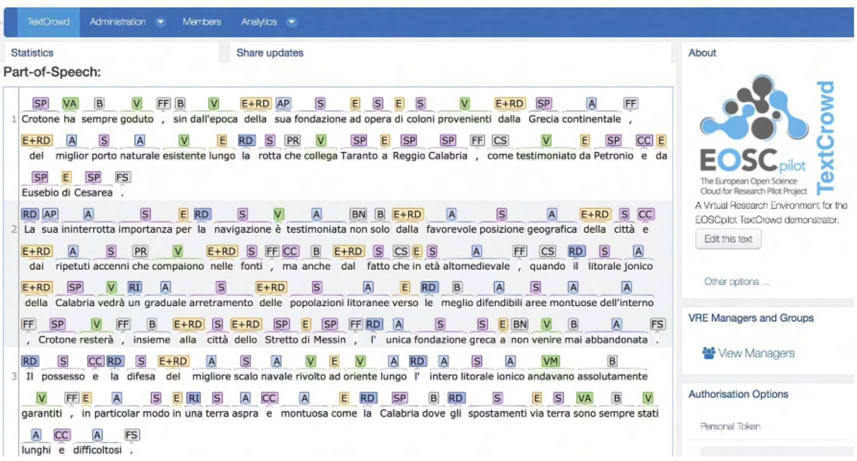
[1] F. Niccolucci et al.: “Managing Full-Text Excavation Data with Semantic Tools”, VAST 2009. The 10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, 2009.

[2] A. Vlachidis et al.: “Excavating Grey Literature: a case study on the rich indexing of archaeological documents via Natural Language Processing techniques and Knowledge Based resources”. ASLIB Proceedings Journal, 2010.

[3] A. Felicetti, F. Murano: “Scripta manent: a CIDOC CRM semiotic reading of ancient texts”, International Journal on Digital Libraries, Springer, 2016.

Please contact:

Achille Felicetti, VAST-LAB, PIN, Università degli Studi di Firenze, Italy, +39 0574 602578
achille.felicetti@pin.unifi.it

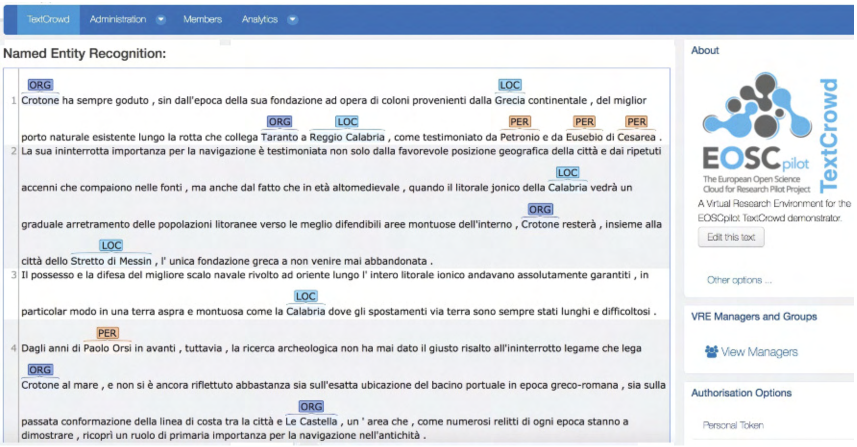


Part-of-Speech:

SP VA B V FF B V E+RD AP S E S E S V E+RD SP A FF
1 Crotone ha sempre goduto , sin dall'epoca della sua fondazione ad opera di coloni provenienti dalla Grecia continentale ,
E+RD A S A V E RD S PR V SP E SP SP FF CS V E SP CC E
del miglior porto naturale esistente lungo la rotta che collega Taranto a Reggio Calabria , come testimoniato da Petronio e da
SP E SP FS
Eusebio di Cesarea .
RD AP A S E RD S V A BN B E+RD A S A E+RD S CC
2 La sua ininterrotta importanza per la navigazione è testimoniata non solo dalla favorevole posizione geografica della città e
E+RD A S PR V E+RD S FF CC B E+RD S CS E S A FF CS RD S A
dai ripetuti accenni che compaiono nelle fonti , ma anche dal fatto che in età altomedievale , quando il litorale jonico
E+RD SP V RI A S E+RD S A E RD B A S A A
della Calabria vedrà un graduale arretramento delle popolazioni litoranee verso le meglio difendibili aree montuose dell'interno
FF SP V FF B E+RD S E+RD SP E SP FF RD A S S E BN V B A FS
, Crotone resterà , insieme alla città dello Stretto di Messin , l' unica fondazione greca a non venire mai abbandonata .
RD S CC RD S E+RD A S A V E V A RD A S A VM B
3 Il possesso e la difesa del migliore scalo navale rivolto ad oriente lungo l' intero litorale ionico andavano assolutamente
V FF E A S E RI S A CC A E RD SP B RD S E S VA B V
garantiti , in particolar modo in una terra aspra e montuosa come la Calabria dove gli spostamenti via terra sono sempre stati
A CC A FS
lunghi e difficoltosi .

A Virtual Research Environment for the EOSCpilot TextCrowd demonstrator.

Figure 1: Part-of-speech recognition in TEXTCROWD using OpenNER framework.



Named Entity Recognition:

ORG LOC
1 Crotone ha sempre goduto , sin dall'epoca della sua fondazione ad opera di coloni provenienti dalla Grecia continentale , del miglior
ORG LOC PER PER PER
porto naturale esistente lungo la rotta che collega Taranto a Reggio Calabria , come testimoniato da Petronio e da Eusebio di Cesarea .
2 La sua ininterrotta importanza per la navigazione è testimoniata non solo dalla favorevole posizione geografica della città e dai ripetuti
LOC
accenni che compaiono nelle fonti , ma anche dal fatto che in età altomedievale , quando il litorale jonico della Calabria vedrà un
ORG
graduale arretramento delle popolazioni litoranee verso le meglio difendibili aree montuose dell'interno , Crotone resterà , insieme alla
LOC
città dello Stretto di Messin , l' unica fondazione greca a non venire mai abbandonata .
3 Il possesso e la difesa del migliore scalo navale rivolto ad oriente lungo l' intero litorale ionico andavano assolutamente garantiti , in
LOC
particolar modo in una terra aspra e montuosa come la Calabria dove gli spostamenti via terra sono sempre stati lunghi e difficoltosi .
PER
4 Dagli anni di Paolo Orsi in avanti , tuttavia , la ricerca archeologica non ha mai dato il giusto risalto all'ininterrotto legame che lega
ORG
Crotone al mare , e non si è ancora riflettuto abbastanza sia sull'esatta ubicazione del bacino portuale in epoca greco-romana , sia sulla
ORG
passata conformazione della linea di costa tra la città e Le Castella , un ' area che , come numerosi relitti di ogni epoca stanno a
dimostrare , ricoprì un ruolo di primaria importanza per la navigazione nell'antichità .

A Virtual Research Environment for the EOSCpilot TextCrowd demonstrator.

Figure 2: Named entities resolution in TEXTCROWD using DBPedia and BabelNet resources.

MTAS – Extending Solr into a Scalable Search Solution and Analysis Tool on Multi-Tier Annotated Text

by Matthijs Brouwer (Meertens Institute, KNAW)

To deliver searchability on huge amounts of textual resources and associated metadata, the Lucene based Apache Solr index provides a well proven and scalable solution. Within the field of Humanities, textual data is often enriched with structures like part-of-speech, named entities, sentences or paragraphs. To include and use these annotations in search conditions and results, we developed a plugin that extends existing Solr functionalities with search and analysis options on these layers.

The Lucene approach to process textual data is based upon tokenisation of the provided information: for each document and field, the occurring tokens or words are stored within the index together with positional information and an optional payload. This offers quick retrieval of matching documents when searching for specific words, and specific sequences can be found by efficiently exploiting positional information on only those documents that contain all words involved. Mtas extends this technique by encoding a prefix and postfix within the value of each token. Since Lucene allows the use of multiple tokens on the same position, this provides the capability to store multiple layers of single positioned tokens, where layers can be distinguished by the applied prefix. Furthermore, to process non-single positioned and hierarchical related elements, additional information is stored within the payload, allowing these items to be stored as single positioned tokens on their first occurring position in the Lucene index structure. Finally, to enable fast retrieval of information based on position or hierarchical relation within the document, forward indices are created.

Although the default Lucene query mechanism can still be applied, specific methods are needed and made available within Mtas to efficiently use the additional encoded information and indices. Based on these methods, within Mtas a parser is provided for the Corpus Query Language (CQL), making it possible for users to define advanced conditions on the annotated text directly in the Solr search request. Since existing Solr functionality is maintained, this can be combined with regular defined conditions on metadata fields within the same document. For selected documents, Mtas can efficiently generate frequency lists for occurring layers, provide statistics over matches for possibly multiple user defined CQL queries, categorize these by one or multiple metadata fields, produce keyword-in-context representations, and group results over one or multiple layers. Mtas fully supports the distributed search capabilities from Solr, providing not only scalability but also making the process of updating and extending the data with new collections easier.

Parsers are available for multiple often XML based annotated document types, e.g. FoLiA, ISO-TEI and CHAT. The mapping by the parser of these usually

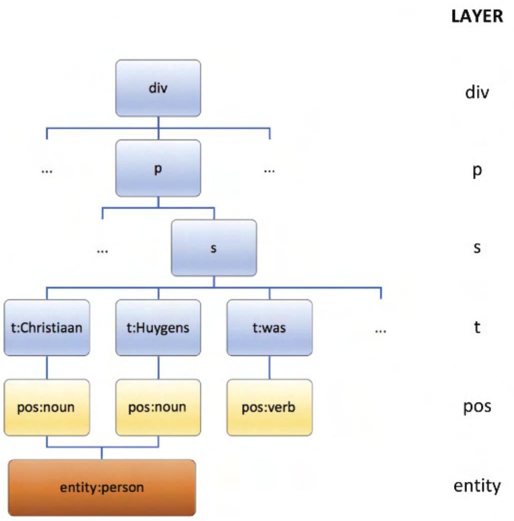


Figure 1: Typical structure of textual data with relations and annotations on multiple tiers as can be processed by Mtas.

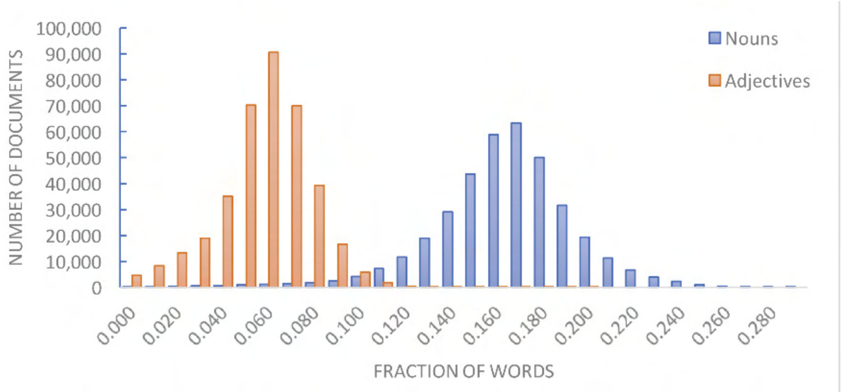
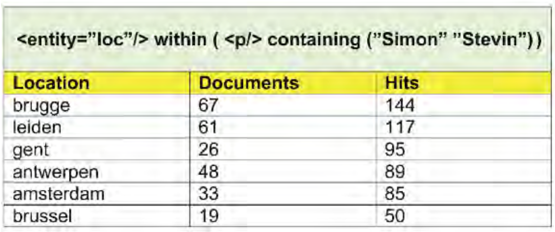


Figure 2: Distribution of the fraction of nouns and adjectives over all documents in the Nederlab collection with at least 500 part-of-speech annotated words.



<entity="loc"/> within (<p/> containing ("Simon" "Stevin"))

Location	Documents	Hits
brugge	67	144
leiden	61	117
gent	26	95
antwerpen	48	89
amsterdam	33	85
brussel	19	50

Figure 3: Location entities in paragraphs containing 'Simon Stevin', constructed with Mtas by grouping all matches on a CQL expression.

somewhat loosely defined formats onto the index can be configured in detail to include in a coherent way (only) those layers that are of interest. Multiple documents and multiple mapping configurations can be combined within the same index and field. Source code, documentation, example configurations and a Docker demonstration version are available from GitHub [L1].

The Nederlab project [L2] is one of the primary use cases for the system. It aims to bring together all digitized texts relevant to the national heritage of the Netherlands, the history of Dutch language and culture (c. 800 – present) in one user-friendly and tool-enriched open access web interface, allowing scholars to simultaneously search and analyse data from texts spanning the full recorded history of the Netherlands, its language and culture. It currently provides access to over 15 million items or documents, containing almost 10 billion positions or words and over 35 billion annotations. This data is distributed over 23 different cores with a total index size from over one terabyte, and hosted on a single Xeon E5 server with 128 GB memory. Query time, of course depending on the type of request, is usually in the order of seconds and limited by server configuration to three minutes.

Some experiments on the use of topic modelling techniques with results from Mtas have been performed. This was heavily based on the expensive generation of frequency lists, and subsequent outcomes are not used to refine or extend analysis and search within Mtas. We would like to improve the efficiency of applying these techniques and offer methods to use results again for further research within conditions on the Solr index. Another ambition involves adjustments of the Mtas index structure to incorporate frequency lists on multiple layers (e.g., to create a list of all used adjectives), which could also improve performance for certain queries. Furthermore, we are interested in implementing an additional query language to explore the stored hierarchical structure, since this is currently not covered very well by CQL. Finally, to better support and integrate sequence-based techniques, it will probably be necessary to extend the index structure with n-grams, whilst also improving certain types of already supported queries.

Links:
[L1] https://kwz.me/hmd
[L2] https://kwz.me/hmc

Reference:

[1] M. Brouwer, H. Brugman, M. Kemps-Snijders, MTAS: A Solr/Lucene based Multi-Tier Annotation Search solution, Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence

Please contact:

Matthijs Brouwer
Meertens Institute, KNAW,
Amsterdam, The Netherlands
matthijs.brouwer@meertens.knaw.nl

Phonetic Search in Audio and Video Recordings

by Ioannis Dologlou and Stelios Bakamidis (RC ATHENA)

A new system uses advanced speech recognition technology to easily and efficiently retrieve information from audio/video recordings just by using keywords.

The massive amount of information produced by today's media (radio, television, etc.) and telecommunications (fixed, mobile telephony, satellite communications, etc) necessitates the use of automatic management strategies. Useful information can be retrieved from audio/video files by using keywords, in the same way as for text files, with a system that automatically searches for appropriate information in audio/video files using a state-of-the-art voice recognition engine. This enables valuable information in broadcast news or telephone conversations to be retrieved easily, quickly and accurately.

This system was developed by Voice-In SA, a spin-off company of the Greek Research Centre RC ATHENA [L1]. The research started in 2008 and the first system was delivered two years later. Research is ongoing to improve the performance and speed of the algorithms involved.

The proposed system implements the most advanced speech recognition technology (large vocabulary, continuous speech, speaker independent). It converts the statistical models of the speech recognition system and adapts them to increase both flexibility and efficiency over the handling of information which is provided by the keywords. In addition the new approach comprises a scoring algorithm for automatic detection of words or phrases that are closest to the user's query.

The system consists of two subsystems. The first subsystem performs a pre-processing on each new archiving material (recordings or video files), so that a file with specific information is created. The second subsystem is the actual core of the system that implements the new algorithms for search and retrieval, simultaneously exploiting the previously stored information.

The input to the system is audio or video files along with some keywords that the user wants to locate in these files. Following a very fast processing of the input data, the system provides information on whether the keywords are present in those files or not. If the outcome of the search is positive, the specific audio or video spots that have been found are mined and supplied to the user accompanied by the exact timing information and their confidence level.

The major advantage of the new approach that makes it unbeatable compared to existing solutions is its ability to fully operate on any subject without any prior learning phase. Consequently it requires no overhead for customisation and/or installation and maintenance. More precisely, the system does not use any lexicon or database that will become outdated and need updating. Furthermore, it can cope with all kinds of words and terminology regardless of their frequency of use. The system has a very user-friendly interface with a comprehensive menu even for non-professionals. It can process all of the following formats: wav, wmv, wma, mp3, mpg, asf and avi.

The new system is useful for several applications in various domains and activities including:

- Automated registration, classification, indexing and efficient, fast and inexpensive recovery of information from audiovisual media.
- Easy access to information produced by state institutions (parliament, government departments, municipalities, communities, etc.).
- Information retrieval from audiovisual material from meetings and general board meetings, corporate bodies etc.
- Forensic applications, i.e., helping to locate people suspected of being involved in illegal activities through automatic monitoring of telephone calls, video recordings, etc.
- Automatic monitoring of air and maritime frequencies in real time to detect incidents, such as mayday, for a prompt response.

Future activities

Future plans focus on the performance of the algorithms both in terms of accuracy and speed. Improving the accuracy involves the creation of a better speech recognition system with a large variety of acoustic models for many different environments (noisy, cocktail party effect etc). Faster search algorithms are also needed for handling the stored information which is created by the first subsystem.

Link:
[L1] https://kwz.me/hmy

References:
[1] S. Vijayarani, A. Sakila: “Multimedia Mining Research – An Overview”, IJCGA, Vol. 5, No.1, Jan 2015, 69-77.
[2] R. Pieraccini: “The Voice in the Machine. Building Computers That Understand Speech”, The MIT Press, 2012, ISBN 978-0262016858.
[3] T. Sainath et al.: “Convolutional neural networks for LVCSR”, ICASSP, 2013.

Please contact:
Ioannis Dologlou
RC ATHENA, Greece, +302106875306
ydol@ilsp.gr

KA3: Speech Analytics for Oral History and the Language Sciences

by Joachim Köhler (Fraunhofer IAIS), Nikolaus P. Himmelmann (Universität zu Köln) and Almut Leh (FernUniversität in Hagen)

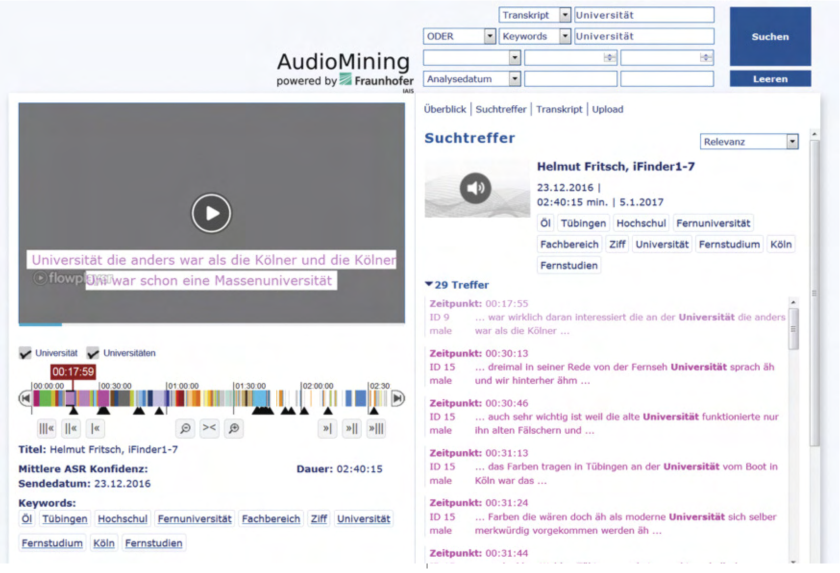
In the project KA3 (Cologne Centre for Analysis and Archiving of Audio-Visual Data), advanced speech technologies are developed and provided to enhance the process of indexing and analysis of speech recordings from the oral history domain and the language sciences. These technologies will be provided as a central service to support researchers in the digital humanities to exploit spoken content.

AAudio-visual and multimodal data are of increasing interest in digital humanities research. Currently most of the data analytics tools in digital humanities are purely text-based. In the context of a German research program to establish centres for digital humanities research, the focus of the three year BMBF project KA3 [L1] is to investigate, develop and provide tools to analyse huge amounts of audio-visual data resources using advanced speech technologies. These tools should enhance the process of transcribing audio-visual recordings semi-automatically and to provide additional speech related analysis features. In current humanities research most of the work is performed completely manually by labelling and annotating speech recordings. The huge effort in terms of time and human resources required to do this severely limits the possibilities of properly including multimedia data in humanities research.

In KA3 two challenging application scenarios are defined. First, in the interaction scenario, linguists investigate the structure of conversational interactions. This includes the analysis of turn taking, back channelling and other aspects of the coordination involved in smooth conversation. A particular focus is on cross-linguitic comparison in order to delimit universal infrastructure for communication from language-specific aspects which are defined by cultural norms. The second scenario is targeting the oral history domain. This research direction uses extended interviews to investigate historical, social and cultural issues. After the recording process the interviews are transcribed manually turning the oral source into a text document for further analysis [1] [L2].

To apply new approaches to analyse audio-visual recordings in these digital humanities research scenarios, Fraunhofer IAIS provides tools and expertise for automatic speech recognition and speech analytics [2] [L3]. For the automatic segmentation and transcription of speech recordings, the Fraunhofer IAIS Audio Mining System is applied and adapted. The following figure shows the user interface of a processed oral history recording.

The recording is transcribed with a large vocabulary speech recognition system based on Kaldi technology. All recognised words and corresponding



Transkript: Universität

ODER Keywords: Universität

Universität die anders war als die Kölner und die Kölner
Uni war schon eine Massenuniversität

Titel: Helmut Fritsch, iFinder1-7

Mittlere ASR Konfidenz:

Dauer: 02:40:15

Sendedatum: 23.12.2016

Keywords:
Öl | Tübingen | Hochschul | Fernuniversität | Fachbereich | Ziff | Universität | Fernstudium | Köln | Fernstudien

Suchtreffer

Helmut Fritsch, iFinder1-7

23.12.2016 |
02:40:15 min. | 5.1.2017

Öl | Tübingen | Hochschul | Fernuniversität | Fachbereich | Ziff | Universität | Fernstudium | Köln | Fernstudien

29 Treffer

Zeitpunkt: 00:17:55
ID 9 ... war wirklich daran interessiert die an der **Universität** die anders
male war als die Kölner ...

Zeitpunkt: 00:30:13
ID 15 ... dreimal in seiner Rede von der Fernseh **Universität** sprach äh
male und wir hinterher ähm ...

Zeitpunkt: 00:30:46
ID 15 ... auch sehr wichtig ist weil die alte **Universität** funktionierte nur
male ihn alten Fälschern und ...

Zeitpunkt: 00:31:13
ID 15 ... das Farben tragen in Tübingen an der **Universität** vom Boot in
male Köln war das ...

Zeitpunkt: 00:31:24
ID 15 ... Farben die wären doch äh als moderne **Universität** sich selber
male merkwürdig vorgekommen werden äh ...

Zeitpunkt: 00:31:44

Figure: Search and Retrieval Interface of the Fraunhofer IAIS Audio Mining System for Oral History.

time codes are indexed in the Solr search engine. Hence, the user can search for relevant query items and directly acquire the snippets of the recognised phrases and the entry points (black triangles) to jump to the position where this item was spoken. Additionally the application provides a list of relevant key words which are calculated by a modified tf-idf algorithm to roughly describe the interview. All metadata is stored in the MPEG-7 format which increases the interoperability with other metadata applications. Mappings to the ELAN format are realised to perform an additional annotation with the ELAN tool. The system is able to process long recordings of oral history interviews (up to three hours). Depending on the recording quality, the initial recognition rate is up to 75 percent. For recordings with low audio quality the recognition process is still quite error-prone. Besides the speech recognition aspect, other speech analytics algorithms are applied. For the interaction scenario, the detection of back-channel effects and overlapping speech segments are carried out.

The selected research scenarios raise challenging new research questions for speech analytics technology. Short back-channel effects are still hard to recognise. On the other hand, speech analytics tools already provide added value in processing huge amounts of new oral history data, thus improving retrieval and interpretation. The collaboration between speech technology scientists and digital humanities researchers is an important aspect of the KA3 project.

Links:

[L1] http://dch.phil-fak.uni-koeln.de/ka3.html
[L2] http://ohda.matrix.msu.edu/
[L3] https://kwz.me/hmH

References:

[1] D. Oard: "Can automatic speech recognition replace manual transcription?", in D. Boyd, S. Cohen, B. Rakerd, & D. Rehberger (Eds.), Oral history in the digital age, Institute of Library and Museum Services, 2012.

[2] C. Schmidt, M. Stadtschnitzer and J. Köhler: "The Fraunhofer IAIS Audio Mining System: Current State and Future Directions", ITG Fachtagung Speech Communication, Paderborn, Germany, September 2016.

Please contact:

Joachim Köhler
Fraunhofer IAIS, Germany
+49 (0) 2241 141900
joachim.koehler@iais.fraunhofer.de

Graph-Based Entity-Oriented Search: Imitating the Human Process of Seeking and Cross Referencing Information

by José Devezas and Sérgio Nunes (INESC TEC and FEUP)

In an information society, people expect to find answers to their questions quickly and with little effort. Sometimes, these answers are locked within textual documents, which often require a manual analysis, after being retrieved from the web using search engines. At FEUP InfoLab, we are researching graph-based models to index combined data (text and knowledge), with the goal of improving entity-oriented search effectiveness.

We live in a world where an ever-growing web is able to deliver a large body of knowledge to virtually anyone with an internet connection. At the same time, the high availability of content has morphed human information seeking behaviour [1]. People expect to find answers to their questions quickly and with little effort. The quality of the answers is frequently tied with the search engine's ability to understand query intent, using information from curated knowledge bases to provide direct answers, based on identified entities and relations, alongside the traditional textual document results. Search engines are greatly dependent on the inverted index, inspired by the back-of-the-book index of printed manuscripts, to rank documents with matching keywords, but they are also increasingly

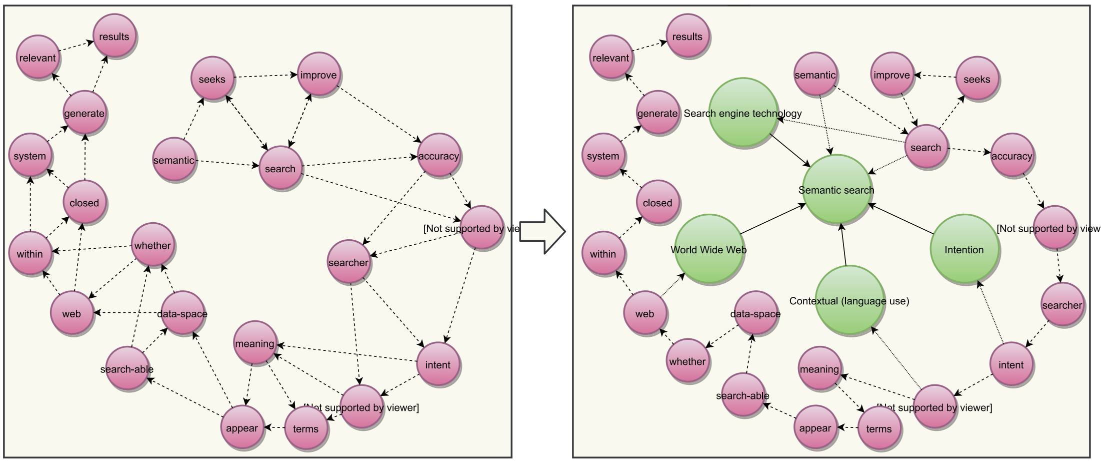


Figure 1: The left side shows the graph-of-word representation for an example document (the first sentence of the Wikipedia page for "Semantic Search"). Each term links to the following two terms, as a way to use indegree to establish the context of a word. The right side shows the proposed graph-of-entity, linking consecutive terms, terms occurring within entity names, and related entities, as a way to unify text and knowledge.

dependent on knowledge bases. While there are automatic methods for knowledge base construction, most search engines still depend on manual curation for this task. On one side, there is the error associated with automatic knowledge base construction and, on the other side, there is the time constraint and domain expertise of manually curating a knowledge base. We propose an intermediate solution based on a novel graph-based indexing structure, with the goal of combining the power of the inverted index with any available and trustworthy information, through established knowledge bases.

Our current research is focused on finding novel ways of integrating text and information through a graph, without losing the properties of terms in an inverted index, and entities and relations in a triple store. The goal is to be able to retain the characteristics of text and knowledge, while combining them, through a unified representation, to improve retrieval. The hypothesis is that by establishing potentially weak links between text and entities, we might be able to support and imitate the human process of seeking and cross referencing information: knowledge-supported keyword-based search. As humans, each of us compiles knowledge from multiple sources (e.g., the world, books, other people), establishing relations between entities and continuously correcting for consistency, based on concurrent information and its trustworthiness. When we have an information need, we either ask someone or consume some sort of media (e.g., a book, a video, an audio lesson) to obtain answers. Let us take, for instance, the task of searching within a book to solve an information need. Specifically, let us assume a back-of-the-book index search for a given set of terms (analogous to the traditional keyword query). Let us then assume that we skip to a page indicated by one of those terms and read a textual passage. How do we determine whether or not it is relevant for our information need? While we already know it contains one of the terms we seek, we must use existing knowledge (ours or otherwise) to assess the relevance of the text.

There is an obvious connection between text and knowledge that isn't being captured by existing search technologies. While there is a clear and growing integration of text-based search and entity-based decorations (e.g., an infobox about the most relevant entity, or a list of entities for the given entity type expressed by the query), the inverted index still exists separately from the knowledge base and vice-versa. Our goal is to explore the opportunity of improving retrieval effectiveness based on a seamless integration of text and knowledge through a common data model, while proposing one or several unifying ranking functions that only decide based on the maximum available information.

We have based our work on the graph-of-word [2], a document representation and retrieval model that defies the term independence assumption of the traditional bag-of-words approach used in inverted indexing. Figure 1 (left) shows the graph-of-word representation for an example document (the first sentence of the Wikipedia page for "Semantic Search"). Each term links to the following two terms, as a way to use indegree (the number of incoming links) to establish the context of a word. We propose the graph-of-entity (Figure 1; right), where we link each term (in pink) only to the following term (dashed line), but also include entity nodes (in green), basic "contained_in" edges between term and entity nodes (dotted line; weak relation based on substring matching), and edges between entity nodes (solid line), representing relations between entities in a knowledge base or, in this case, indirectly based on the hyperlinks for the Wikipedia article. The objective is to unify text and knowledge retrieval as a combined task, in order to use structured and unstructured data to provide better answers for the information needs of the users.

Links:

[L1] http://infolab.fe.up.pt
[L2] http://ant.fe.up.pt

References:

[1] S A Knight and Amanda Spink: "Toward a web search information behavior model", in Web Search, pages 209-234. Springer, 2008.

[2] F. Rousseau, M. Vazirgiannis: "Graph-of-word and TW-IDF: new approach to ad hoc IR", in Proc. of the 22nd ACM international conference on Information & Knowledge Management, pp 59-68. ACM, 2013.

Please contact:

José Devezas, Sérgio Nunes, INESC TEC and Universidade do Porto
jld@fe.up.pt, ssn@fe.up.pt
http://josedevezas.com

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

ContentCheck: Content Management Techniques and Tools for Fact-checking

by Ioana Manolescu (Inria Saclay and Ecole Polytechnique, France)

Data journalism and journalistic fact-checking make up a vibrant area of applied research. Content management models and algorithms have the potential to tremendously multiply their power, and have already started to do so.

The immense value of big data has recently been acknowledged by the media industry, with the coining of the term "data journalism" to refer to journalistic work inspired by data sources. While data is a natural ingredient of all reporting, the increasing volumes of available digital data as well as its increasing complexity lead to a qualitative jump, where technical skills for working with data are stringently needed in journalism teams.

An ongoing collaborative research programme focused on novel content management techniques applied to data journalism and fact-checking has recently been initiated by: Inria, the LIMSI lab (Université Paris Saclay, CNRS and Université Paris Sud), Université Rennes 1, Université Lyon 1 and the "Les Décodeurs" fact-checking team of Le Monde, France's leading newspaper [L1]. Here, content is broadly interpreted to denote structured data, text, and knowledge bases, as well as information describing the social context of data being produced and exchanged between various actors. The project, called ContentCheck [L2], is sponsored by ANR, the French National Research Agency, and is set to run until 2019.

The project goals are twofold:

- First, in an area rich with sensational fake news, debunked myths, social media rumors, and ambitious start-ups, we aim at a comprehensive analysis of the areas of computer science from which data journalism and fact-checking can draw ideas and techniques. Questions we seek to answer include: what kinds of content are involved? What is their life cycle? What types of processing are frequently applied (or needed!) in journalistic endeavours? Can we identify a blueprint architecture for the ideal journalistic content management system (JCMS)?
- Second, we seek to advance and improve the technical tools available for such journalistic tasks, by proposing specific high-level models, languages, and algorithms applied to data of interest to journalists.

Prior to the start of the project, interviews with mainstream media journalists from Le Monde, The Washington Post and the Financial Times have highlighted severe limitations of their JCMSs These are typically restricted to archiving published articles, and providing full-text or category-based searches on them. No support is available for storing or processing external data that the journalists work with on a daily basis; newsrooms rely on ad-hoc tools such as shared documents and repositories on the web, and copied files to and fro as soon as processing was required. The overhead, lost productivity, privacy and reliability weaknesses of this approach are readily evident.

The main findings made in our project to date include:

- Data journalism and fact-checking involve a wide range of content management and processing tasks, as well as human-intensive tasks performed individually (e.g., a journalist or an external expert of a given field whose input is solicited – For instance, ClimateFeedback is an effort to analyse media articles about climate change by climate scientists. See, for example, [L3]), or collectively (e.g., readers can help flag fake news in a crowd-sourcing scenario, while a large consortium of journalists may work on a large news story with international implications. The International Consortium of Investigative Journalism is at the origin of the Panama Papers [L4] dis-

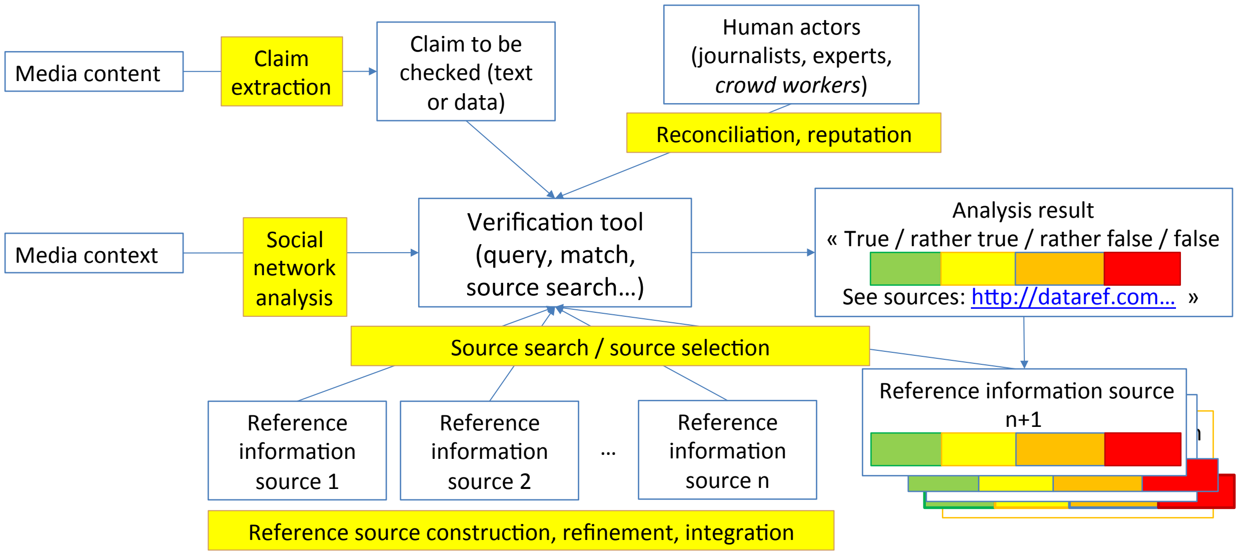

Media content
Claim extraction
Claim to be checked (text or data)
Human actors (journalists, experts, crowd workers)
Reconciliation, reputation
Media context
Social network analysis
Verification tool (query, match, source search…)
Analysis result
« True / rather true / rather false / false
See sources: http://dataref.com… »
Source search / source selection
Reference information source 1
Reference information source 2
…
Reference information source n
Reference information source n+1
Reference source construction, refinement, integration

Figure 1: Flow of information in fact-checking tasks, and relevant research problem from the content/data/ information management area.

closure concerning tax avoidance through tax havens).
- Content management tasks include the usual CRUD (create, read, update, delete) data cycle, applied both to static content (e.g., web articles or government PDF reports) and dynamic content such as provided by social media streams.
- Stream monitoring and stream filtering tools are highly desirable, as journalists need help to identify, in the daily avalanche of online information, the subset worth focusing on for further analysis.
- Time information attached to all forms of content is highly valuable. It is important to keep track of the time-changing roles (e.g., elected positions) held by public figures, and also to record the time when statements were made, and (when applicable) the time such statements were referring to (e.g., when was a certain politician's spouse employed, and when did the politician share this information).

Work to design a single unified architecture for a content management tool dedicated to fact-checking is still ongoing as we write. The overall vision we currently base our analysis on is outlined in Figure 1. Claims are made through various media, and (importantly) in a context, in which one can find the claim's authors, their institutions, friend and organisational affiliations etc. Claims are fact-checked against some reference information, typically supplied by trustworthy institutions, such as statistics national institutes (INSEE in France, the Office for National Statistics in the UK) or trusted experts, such as well-established scientists working on a specific topic. Claims are checked by human users (journalists, scientists, or concerned citizens), possibly with the help of some automated tools. The output of a fact-checking task is a claim analysis, which states parts of the claims that are true, mostly true, mostly false etc., together with references to the trustworthy sources used for the check. Fact-checking outputs, then, can be archived and used as further reference sources.

Scientific outputs of the project so far include a light-weight data integration platform for data journalism [1], an analysis of EU Parliament votes highlighting unusual (unexpected) correlations between the voting patterns of different political groups [2], and a linked open data extractor out of INSEE spreadsheets [3]. Our project website is available at: [L2].

Links:
[L1] https://kwz.me/hmE
[L2] https://kwz.me/hmK
[L3] https://kwz.me/hmS
[L4] https://panamapapers.icij.org/

References:

[1] R. Bonaque, et al.: "Mixed-instance querying: a lightweight integration architecture for data journalism" (demonstration), PVLDB Conference, 2016.
[2] A. Belfodil, et al.: "Flash points: Discovering exceptional pairwise behaviors in vote or rating data", ECML-PKDD Conference, 2017.
[3] T. D. Cao, I. Manolescu, X. Tannier: "Extracting Linked Data from statistic spreadsheets", Semantic Big Data Workshop 2017.

Please contact:

Ioana Manolescu, Inria Saclay and Ecole Polytechnique, France
ioana.manolescu@inria.fr

Argumentation Analysis of Engineering Research

by Mihály Héder (MTA-SZTAKI)

Engineering research literature tends to have fewer citations per document than other areas of science. Natural language processing, argumentation structure analysis, and pattern recognition in graphs can help to explain this by providing an understanding of the scientific impact mechanisms between scientists, engineers and society as a whole.

Institutionalised engineering research often needs to follow research policy that was designed with natural science in mind. In this setting, individual academic advancement as well as research funding largely depends on scientific indicators. A dominant way of measuring impact in science is counting citations at different levels and in different dimensions: for individual articles and journals as well as for individuals and groups. If an engineering research group does not deliver on these metrics, it might compensate for this with the revenue it generates. But this strategy increasingly pushes a group from basic engineering research activity to applied and short-term profitable research, since basic research does not result in income.

Our main research question is: why are there fewer citations per document in this field compared to almost any other branch of science? This has direct consequences on the aforementioned scientific metrics of the field. We are also addressing the questions: are conventional citations good indicators for engineering research at all? And what would be the ideal impact measurement mechanism for engineering research?

Our work approaches the problem by investigating both the argumentative structure [1] of engineering research articles and the directed graphs that represent citations between papers. For investigating individual papers we use natural language processing, including named entity recognition, clasterization, classification and keyword analysis. During our work we rely on publicly available Open Access registries and citation databases represented in linked open datasets.

We are testing several initial hypotheses which might explain the low number of citations:
- there are virtually no long debates that manifest themselves in debate-starter papers and follow-ups; (
- The audience itself on which engineering research has impact does not publish in big numbers;
- There are some additional effects: references to standards, design documents and patents are often just

implied and not made explicit – and even when they are, they do not count.

The first hypothesis is tested with graph pattern matching. In this case we are defining an abstract pattern in the citation graphs of noted debates in the fields of philosophy of science and philosophy of technology. Then, we attempt to recognise similar patterns in the engineering research literature.

To test the second hypothesis we look into additional sources, like standards and software code that are known to be applications of certain research papers. Then we investigate the publication history of the creators of those applications, to see if they report those applications in publications. Here we rely on publicly available citation databases and searchable databases provided by big publisher and internet search firms.

For the third hypothesis we have invented the definitions of several types of "implicit citations". Implicit citations are cases where the impact of a research article is clear in some work or artefact – standards, software code, patents, etc – but because of the nature of the work the impact never appears as a citation in any database. A typical example of this is the usage of a particular algorithm in software. While it is not appropriate to consider these kinds of "implicit citations" as equal to citations from within prestigious papers, they still point to the research's effects on industry and society. Public funding is often justified by the advantages a research direction eventually brings to the taxpayer and society, so it is good to have an objective metric.

The preliminary results indicate that the low number of citations in the field of engineering research can, to a great extent, be explained by the hypotheses above. We have also identified other unanticipated factors, namely a proxy effect that lowers the overall number of citations as well as a tendency in the field to cite a well-known named entity (like the name of an algorithm) but not referencing to it in a bibliographically correct way.

Science metrics on which researchers need to deliver are ways of limiting the freedom of inquiry, since they prescribe the publication types researchers need use, as well as the places – prestigious journals, conferences – where they must publish. This is usually done in good will and with the intent of improving the quality of science, but the usual metrics can be detrimental to the cause of an envisioned engineering science [2]. Since the need for some kind of metrics is likely to remain, the project will propose alternatives that measure engineering research impact more inclusively by incorporating design documents, patents, standards, and source code usage.

References:

[1] S. Teufel: The Structure of Scientific Articles: Applications to Citation Indexing and Summarization. Center for the Study of Language and Information-Lecture Notes, 2010.

[2] B. Martin The use of multiple indicators in the assessment of basic research. Scientometrics, 1996, 36.3: 343-362.

Please contact:

Mihály Héder
SZTAKI, Hungarian Academy of Sciences, Hungary
mihaly.heder@sztaki.mta.hu
36 1 279 6027

A Back Bone Thesaurus for Digital Humanities

by Maria Daskalaki and Lida Charami (ICS-FORTH)

In order to integrate new digital technologies and traditional research in the humanities and enable collaboration across the various scientific fields, we have developed a coherent overarching thesaurus with a small number of highly expressive, consistent upper level concepts which can be used as the starting point for harmonising the numerous discipline and even project specific terminologies into a coherent and effective thesaurus federation.

Digital humanities is a relatively new interdisciplinary field which involves the integration of emerging new digital technologies with traditional research in the humanities in order to ensure the long term preservation of knowledge and enable collaboration across the various humanities fields. This integration, however, is not as easy and straight-forward as it sounds as it entails the creation and use of a "common language" in the form of a classification scheme that would enable the communication between different disciplines. The actual state-of-play, however, is somewhat different, with different groups of scholars usually developing their own jargon in order to build thematic vocabularies that are discipline or even application specific. As Barry Smith [1] observes "different databases may use identical labels but with different meanings; alternately the same meaning may be expressed via different names". This inevitably introduces an unnecessary fragmentation of knowledge that inhibits research and collaboration. Given this situation, there is clearly an urgent need to create a common scheme that would enable interoperability between the different scholarly fields and thus support researchers by giving them access to uniformly marked up datasets for query and by providing a guide for the production of systematic terminologies which would avoid methodological errors that typically lead to inconsistencies and incompatibilities between classification systems.

Despite the clear challenges to the construction of such a unifying framework, we argue that "a global knowledge network" [2] is feasible. Building on a concentrated research programme into classification methodology, we have developed a system, the Back Bone Thesaurus (BBT) [L1], that aims to allow access, compatibility and comparison across heterogeneous [3] classification systems.

This system, elaborated after the research of a multi-disciplinary team of experts, is based on a consistent methodology designed to enable intersubjective and interdisciplinary classification development and integration without forcing specialists and experts to abandon their own terminology. The methodology relies on the principle of faceted classification and the idea that a limited number of top-level concepts can become a substantial tool to harmonise the numerous discipline and even project specific terminologies into a coherent and effective federation in which consistency can progressively be carried from the upper layers to the lower ones.

In order to define the BBT facets, we started by examining existing vocabularies from the fields of history, archaeology, ethnology, philosophy of sciences, anthropology, linguistics, theatre studies, musicology and history of art, we analysed these data using a bottom up strategy in order to discover appropriate upper level concepts. The research consciously avoided the projection of any preconceived formulations of knowledge onto the material, precisely in order to identify the broader, fundamental categories that would be applicable across the humanities. The top level concepts thus derived, despite their generality, can be easily specialised in order to express the particular meaning of the different domains without leading to inconsistencies. This is achieved through the detection of the intensional properties of these concepts and the rigorous and proper application of the IsA relationship.

In order to express the exact meaning of the top level terms/concepts defined in the BBT, we provide explicit definitions on the basis of their intensional properties which cannot be replaced without loss of meaning since they are the sum of the properties, state of affairs, qualities that constitute the necessary and sufficient conditions for identifying a term/concept.

The BBT facets are further subdivided into a number of hierarchies using the IsA relation which dictates that the scope of each narrower term subsumed under a broader term must fall completely within the scope of the broader term. In other words, every subsumed term must belong to the same inherent category as its broader concept. Using the IsA relation as the criterion for building the BBT hierarchies ensures that consistency is maintained since all narrower terms must possess all the fundamental properties attributed to the broader concepts of the hierarchy into which they are subsumed. In other words, by using the IsA relation we avoid categorical errors that may result from the subsumption of terms under facets or hierarchies, which have properties different than those of the higher level terms. The strict, proper application of the IsA relation thus serves as a logical control to avoid contradictions and achieve objectivity and interdisciplinarity.

The BBT is an ongoing work and we are currently in the process of reviewing the material we have at our disposal in order to identify additional facets and hierarchies.

Link:
[L1] http://www.backbonethesaurus.eu/

References:

[1] B. Smith: “Ontology”, in The Blackwell Guide to the Philosophy of Computing and Information, L. Floridi, ed. Oxford: Blackwell, 2004, p. 158.

[2] M. Doerr, D. Iorizzo: “The dream of a global knowledge network – A new approach”, in Journal on Computing and Cultural Heritage, 2008, 1(1), p. 1.

[3] M. Daskalaki, M. Doerr: “Philosophical background assumptions in digitized knowledge representation systems”, in Dia-noesis: A Journal of Philosophy, 2017, (3), 17-28.

Please contact:

Maria Daskalaki
ICS-FORTH, Greece
daskalak@ics.forth.gr

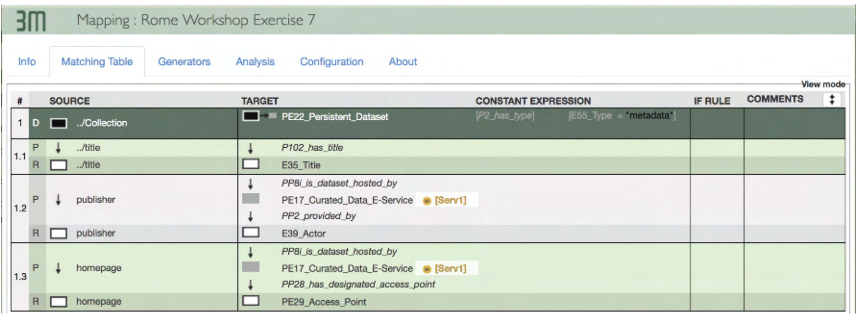
Meeting the Challenges to Reap the Benefits of Semantic Data in Digital Humanities

by George Bruseker, Martin Doerr and Chryssoula Bekiari (ICS-FORTH)

In the era of big data, digital humanities faces the ongoing challenge of formulating a long-term and complete strategy for creating and managing interoperable and accessible datasets to support its research aims. Semantics and formal ontologies, properly understood and implemented, provide a powerful potential solution to this problem. A long-term research programme is contributing to a complete methodology and toolset for managing the semantic data lifecycle.

The field of semantics and the use of formal ontologies for representing research data are key areas of research in the digital humanities, aiming to support the sustainable development of interoperable and accessible datasets for use by research communities in the era of big data. Coined by Berners-Lee, semantic data refers to data which is machine processable and human readable. Formal ontologies provide explicit and disciplined means of producing such data, to ensure its wide compatibility and clear interpretability.

At The Centre for Cultural Informatics (CCI) of ICS-FORTH, based in Heraklion, Greece, we have been researching a comprehensive solution to the complete lifecycle of semantic data use supported by formal ontologies. Our research is nearly at a point where it can be applied within digital humanities by bringing together the basic methods and tools for the distinct steps of this cycle: data modelling, mapping and transformation, querying and management. In particular, research at the CCI focuses on development to fill gaps or improve methodologies in these key steps.



#		SOURCE	TARGET	CONSTANT EXPRESSION	IF RULE	COMMENTS
1	D	../Collection	PE22_Persistent_Dataset	[P2_has_type] [E55_Type = "metadata"]		
1.1	P	../title	*P102_has_title*			
	R	../title	E35_Title			
1.2	P	publisher	*PP8i_is_dataset_hosted_by* PE17_Curated_Data_E-Service [Serv1] *PP2_provided_by*			
	R	publisher	E39_Actor			
1.3	P	homepage	*PP8i_is_dataset_hosted_by* PE17_Curated_Data_E-Service [Serv1] *PP28_has_designated_access_point*			
	R	homepage	PE29_Access_Point			

User interface of X3ML declarative mapping tool.

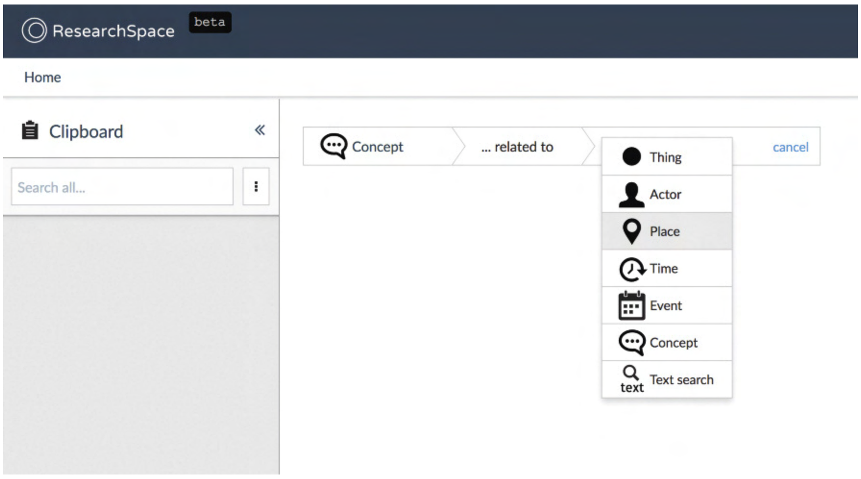


Figure 2: UI of ResearchSpace fundamental categories and relations query tool, ©ResearchSpace.

Before semantic data can be integrated, an adequate model for the domain must be elaborated. Semantic data modelling typically follows one of two basic strategies: the elaboration of complex, all inclusive models or restriction to modelling of a highly focussed domain. Both produce highly useful models, that nevertheless display certain limitations for broad interoperability. The limitations of these strategies tend to lie, on the one hand, in a powerful, general integration with less detailed integration at the leaf level in complex models (e.g.: INSPIRE), and, on the other hand, extremely tight integration of data with lack of relation to the general in more compact models (e.g.: FOAF). As a coordinating member of the international collaborative CIDOC CRM Special Interest Group (SIG), [L1] working under the aegis of the International Council of Museums (ICOM), CCI follows a different approach.

Adopting a bottom up development process that works from actual data structures, the CRM SIG has produced an ontology for cultural heritage and e-sciences which provides the general integrative functions of a base ontology. The base model of CIDOC CRM is currently in the sixth revision of its community version with the fifth revision standing as the base of the last ISO release in 2014 [1]. The long term success in uptake of this model has laid the foundation for community collaboration with experts from various disciplines to create harmonised extensions including: FRBRoo and PRESSoo for library data; CRMdig, CRMinf, and CRMsci which collectively provide provenance data in the respective areas of digitisation, argumentation and observation sciences; and CRMarchaeo and CRMba which support reasoning over archaeological practice. The innovation of this extension development process is the collaborative work with specialist communities to elaborate harmonised extensions to the base model which enable the representation of special domains of research while maintaining compatibility with the top level model.

Having a general ontological framework available to express their data, researchers require a means to translate existing data into the common expression. Development work at CCI has created the X3ML Suite which provides an innovative language, database and data mapping tool for generating completely declarative mappings from any XML data source to any RDFS encoded ontology. This suite of functionalities allows domain specialists to carry out and track mapping processes to the CRM or other suitable ontologies on their own without having to rely on the mediation of computer science specialists. [2] Together with a tool for easily viewing/reviewing RDF data (see article by Minadakis et. al. in the section “Research and Innovation” of this issue), this suite provides a platform for managing large scale semantic mapping processes without restrictions to a specific schema.

Once expressed in a common, but still complex, semantic format, there is still the challenge of how to provide researchers a tool to query this semantic network without necessarily having to learn complex query languages such as SPARQL or the nuances of use of a large ontology. Methodological work at the CCI has produced a theory of Fundamental Categories and Relations which describes how to specify generalist queries over a complex model that will bring back relevant results to researchers by providing an intuitive and semantically consistent abstraction over the complexities of the ontology [3]. This methodology has been taken up and developed as a key tool by the Researchspace project in its development of an open source platform for semantic data in the cultural heritage and digital humanities domains [L2].

Finally, to ensure the sustainable development and use of semantically encoded resources, a complete strategy to the semantic data life cycle must be elaborated. CCI presently participates in the Parthenos project [L3] developing a conceptual data model and architecture for long term semantic data integration and curation, that aims to model the intergration process itself and thereby support long term, on-demand integration tasks and the monitoring thereof.

In order to meet the challenges and take advantage of the benefits of semantic data in Digital Humanities in the era of big data, a complete strategy and set of tools to cover the basic elements of the semantic data lifecycle is essential.

With the maturation of a base model, creation of a declarative mapping tool and language, a generalising query function and a model and method for managing integration processes, we believe that the key elements for meeting these challenges now lie in place.

Links:
[L1] http://www.cidoc-crm.org/
[L2] http://www.researchspace.org/
[L3] http://www.parthenos-project.eu/

References:

[1] ISO: ISO 21127: 2014, Information and documentation – a reference ontology for the interchange of cultural heritage information, 2nd edn., 2014.

[2] N. Minadakis, Y. et al.: "X3ML Framework: An effective suite for supporting data mappings, Extending. Proceedings of the Workshop on Extending, Mapping and Focusing the CRM co-located with 19th International Conference on Theory and Practice of Digital Libraries (2015), Poznań, Poland, September 17, 2015", CEUR Workshop Proceedings 1656, 1-12, 2016.

[3] K. Tzompanaki & M. Doerr: Fundamental Categories and Relationships for Intuitive querying CIDOC-CRM based repositories, Technical Report, 2012.

Please contact:
George Bruseker
Centre for Cultural Informatics
ICS-FORTH, Greece
bruseker@ics.forth.gr

HumaReC: Continuous Data Publishing in the Humanities

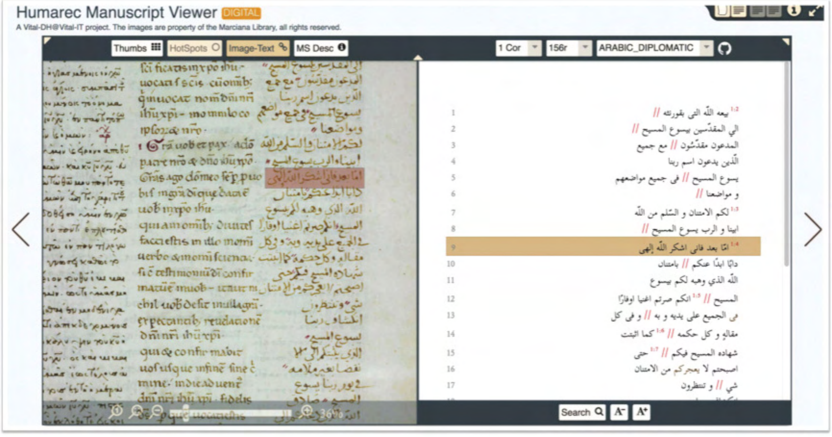
by Claire Clivaz, Sara Schulthess and Anastasia Chasapi (SIB)

HumaReC, a Swiss National Foundation project, aims to test a new model of digital research partnership: from original document source to publisher. Its object of study is a trilingual, 12th century, New Testament manuscript, which is used by HumaReC to test continuous data publishing.

HumaReC [L1] is a Vital-DH@Vital-IT project funded by the Swiss National Foundation. Under the leadership of Claire Clivaz, it started in October 2016 and has been running for two years. The project is based at Vital-IT (Lausanne, CH), under the guidance of group leader Ioannis Xenarios from the Swiss Institute of Bioinformatics. The team is composed of a mixture of humanities and computing scholars; Sara Schulthess is the main researcher.

The aim of HumaReC is to investigate how humanities research is reshaped and transformed by the digital rhythm of data production and publication; it also aims to establish the best practices for Digital Humanities research. As a test-case, the study is focussed on a unique, trilingual New Testament manuscript: Marciana Gr. Z. 11 (379), written in Greek, Latin and Arabic. In the spirit of the OA2020 initiative [L4], continuous data publishing is being tested in partnership with all the research network stakeholders: the Marciana library (Venice, Italy), the Edition Visualization Technology (EVT) [L2], the Transkribus [L3] research teams, and the publisher Brill.

Rhythm is a central notion in the structure of HumaReC and we have chosen this key-concept, based notably on Meschonnic analysis [1], to observe the changes happening in digital humanities research which has been premised on printed culture for a long time. A two to three year research project in humanities has traditionally been characterised by the writing, editing and publication of a final, printed book, often delayed to a certain date after the end of the project. This delay was even considered proof of authentic, high-level research in humanities, certified by an established book series. The digital transition is creating a completely new research paradigm in part due to the publishing of formats such as videos, short messages or draft papers, social media and blogs, all before the research is even completed and peer-reviewed. How is it possible to develop certified and continuous data publishing digital research in the humanities?


Humarec Manuscript Viewer
DIGITAL
A Vital-DH@Vital-IT project. The images are property of the Marciana Library, all rights reserved.
Thumbs
HotSpots
Image-Text
MS Desc
1 Cor
156r
ARABIC_DIPLOMATIC
Search
A⁻
A⁺

1:2 بيعه اللّه التي بقورنثه //
الي المقدّسين بيسوع المسيح //
المدعون مقدّسون // مع جميع
الّذين يدعون اسم ربنا
يسوع المسيح // فى جميع مواضعهم
و مواضعنا //
1:3 لكم الامتنان و السّلم من اللّه
ابينا و الرب يسوع المسيح //
1:4 امّا بعد فانى اشكر اللّه إلهي
دائمًا ابدًا عنكم // بامتنان
اللّه الذي وهبه لكم بيسوع
المسيح // 1:5 انكم صرتم اغنيا اوفارًا
في الجميع على يديه و به // و فى كل
مقالهٍ و كل حكمه // 1:6 كما اثبتت
شهاده المسيح فيكم // 1:7 حتى
اصبحتم لا يعجزكم من الامتنان
شي // و تنتظرون

Figure 1: Manuscript viewer f. 156r © Marciana Library all rights reserved.

As the project's first step, a virtual research environment was created for HumaReC, allowing the research process and results to be made continuously available. It provides a manuscript viewer in fully open access that includes quality images of the manuscript and three columns of transcriptions (see Illustrations below). The manuscript viewer is based on EVT open source technology and is a development of a previous project's viewer [2]. The improved viewer offers additional features such as linking between text and image. In addition to this, an annotation system will allow users to directly comment on the manuscript viewer. Secondly, Transkribus, a

Handwritten Text Recognition tool, will also be tested by HumaReC: a certain number of words transcribed by hand allows a learning machine to be trained to recognise specific writing in a manuscript. The results are forthcoming: we will be able to compare the time taken by purely hand transcription with results produced by a human/machine team.
Finally, three publication formats were chosen for the continuous dissemination of the data:

- The virtual research environment itself; it received an ISSN (2504-5075) from the Swiss National Library: all the published material associated with the project can be referred to with this number. An international editorial board is providing project feedback and input on its research results.
- The research blog is an important interactive continuous publishing process. We regularly update the blog about the development of the project and the research results. We also encourage discussions by being present on social media (Facebook and Twitter).
- The web-book, continuously written in open access, summarises the research in a long, structured text, similar to a conventional monograph but related to the data. It is produced in partnership with the publisher Brill. At the end of the project, it will be peer-reviewed, and hopefully published with its own ISSN by Brill.

We are confident that HumaReC will establish a new research and publishing model, including potential commercial developments for all interested publishers.

Acknowledgement
Many thanks to Harley Edwards for his English proof-reading, to the Marciana library for the image copyright, to the reviewers for their useful remarks, and to Martial Sankar and Ioannis Xenarios for their collaboration and support in our Vital-DH projects@Vital-IT (SIB).

Links:
[L1] https://humarec.org;
http://p3.snf.ch/project-169869
[L2] http://evt.labcd.unipi.it/
[L3] http://transkribus.eu
[L4] http://oa2020.org

References:
[1] C. Clivaz et al.: “Editing New Testament Arabic Manuscripts in a TEI-base: fostering close reading in Digital Humanities”, JDMDH 3700, 2017.
[2] H. Meschonnic: “Critique du rythme. Anthropologie historique du langage”, 1982.

Please contact:
Claire Clivaz, Swiss Institute of Bioinformatics, Switzerland,
+41216924060, claire.clivaz@sib.swiss

A Data Management Plan for Digital Humanities: the PARTHENOS Model

by Sheena Bassett (PIN Scrl), Sara Di Giorgio (MIBACT-ICCU) and Paola Ronzino (PIN Scrl)

Understanding how data has been created and under which conditions it can be reused is a significant step towards the realisation of open science.

PARTHENOS [L1] is a Horizon 2020 project funded by the European Commission that aims at strengthening the cohesion of research in the broad sector of linguistic studies, cultural heritage, history, archaeology and related fields through a thematic cluster of European research infrastructures. PARTHENOS is building a cross-disciplinary virtual environment to enable researchers in the humanities to have access to data, tools and services based on common policies, guidelines and standards. The project is built around two European Research Infrastructure Consortia (ERICs) from the humanities and arts sector: DARIAH [L2] (research based on digital humanities) and CLARIN [L3] (research based on language data), along with ARIADNE [L4] (digital archaeological research infrastructure), EHRI [L5] (European Holocaust research infrastructure), CENDARI [L6] (digital research infrastructure for historical research), CHARISMA [L7] and IPERION-CH [L8] (EU projects on heritage science) and involves all the relevant integrating activities projects.

Since 2016, the Horizon 2020 Programme has produced guidelines on FAIR data management [L9] to help Horizon 2020 beneficiaries make their research data findable, accessible, interoperable and reusable (FAIR) [L10]. Funded projects are requested to deliver an implementation of DMP which aims to improve and maximise access to and reuse of research data generated by the projects. This is in line with Commission’s policy actions on open science to reinforce the EU’s political priority of fostering knowledge circulation. Open science is in practice about “sharing knowledge as early as practically possible in the discovery process” and because DMPs gather information about what data will be created and how, and outline the plans for sharing and preservation, specifying the nature of the data and any restrictions that may need to be applied, these plans ensure that research data is secure and well-maintained during a project and beyond, when it might be shared with others. DMPs are key elements to knowledge discovery and innovation and to subsequent data and knowledge integration and reuse.

Special attention has been paid to the development of a PARTHENOS data management plan which builds on the Horizon2020 DMP template. This has resulted in a template (draft) which aims to address the domain-specific procedures and practices within the humanities, taking into consideration standards and guidelines used in data management that are relevant for PARTHENOS specific research communities, which includes archaeologists, historians, linguists, librarians, archivists, and social scientists.

The PARTHENOS DMP template has been enriched and tailored with specifications from the humanities which were derived from a survey carried out among the consortium's experts. To gather these specifications, representatives of the PARTHENOS communities were asked to describe their daily data management procedures in detail. The questionnaire was structured according to the various phases of the data life cycle and was then mapped to the FAIR principles. Each respondent had the opportunity to choose his/her role (e.g., researcher creating data / repository provider) and to provide insight into their best practices through the form.

The PARTHENOS DMP template will satisfy different stakeholders, such as institutional repositories, funded projects and researchers each of which have an individual perspective on the data quality and FAIRness issues. The PARTHENOS DMP template will be divided into three different levels: a first level which includes a set of core general requirements irrespective of discipline, a second level including domain specific requirements and a third level which is project-based. To help users in completing a data management plan, a set of guiding statements for the specific disciplines will be provided.

At present, the PARTHENOS DMP template collects the high-level requirements that satisfy each community of researchers involved in the project, with a list of recommended answers that will support the compilation of the DMP. A second stage of this work will concern the production of PARTHENOS community-specific DMP templates, which will be included in the final version of the "Guidelines for common policies implementation".

Further work will concern the creation of a DMP template addressing institutions that manage repositories. Since enabling interoperability is a great benefit for researchers, repository providers should be able to explain in depth how to provide data to them in the best way. Through the envisaged template, PARTHENOS will provide the right tools to repository providers to be able to offer standardised answers and guidance, and to liaise with researchers that are looking to deposit their data with them.

Thanks to the PARTHENOS DMP template, researchers will be able to freely access, mine, exploit, reproduce and disseminate their data and identify the tools needed to use the raw data for validating research results, or provide the tools themselves, a significant step towards to the realisation of open science.

Visit the PARTHENOS website [L1] for more information and subscribe to the newsletter to keep up to date with developments.

Links:

[L1] www.parthenos-project.eu
[L2] www.dariah.eu/
[L3] www.clarin.eu/
[L4] www.ariadne-infrastructure.eu
[L5] www.ehri-project.eu/
[L6] www.cendari.eu/
[L7] kwz.me/hTk
[L8] www.iperionch.eu/
[L9] kwz.me/hTO
[L10] kwz.me/hTo

References:

[1] The Three Os – Open innovation, open science, open to the world – a vision for Europe, produced by the European Commission's Directorate-General for Research & Innovation, 2016-05-17, https://kwz.me/hm4

[2] Wilkinson, M. D. et al: "The FAIR Guiding Principles for scientific data management and stewardship", Sci. Data3:160018 doi: 10.1038/sdata.2016.18 (2016).

Please contact:

Sheena Bassett, PIN Scrl, Italy
sheena.giess@gmail.com

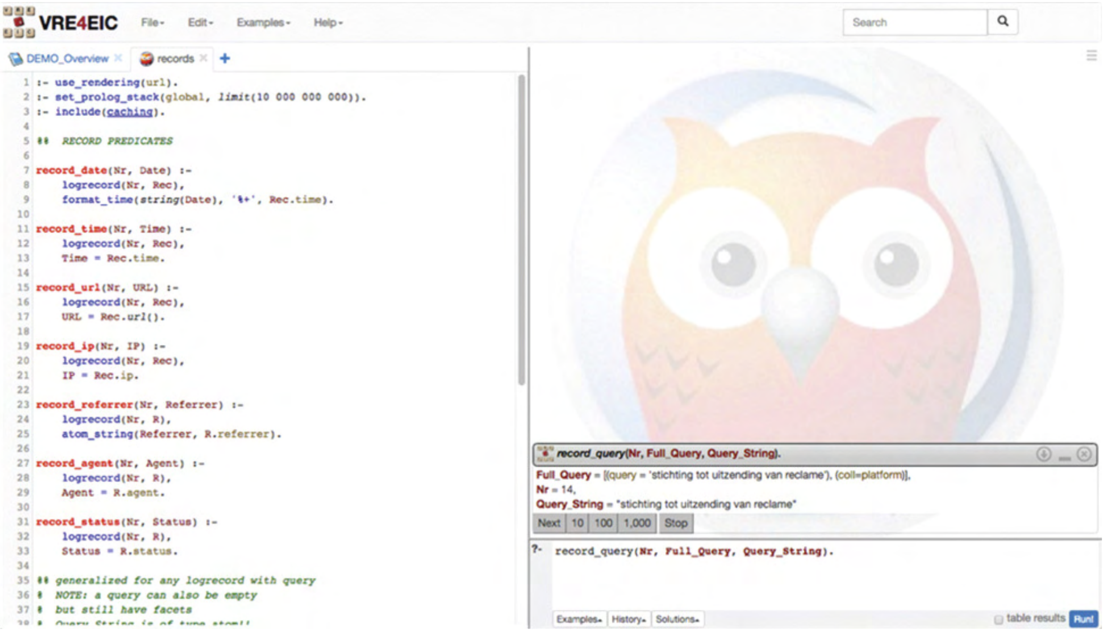
Trusting Computation in Digital Humanities Research

by Jacco van Ossenbruggen (CWI)

Research in the humanities typically involves studying specific and potentially subjective interpretations of (historical) sources, whilst the computational tools and techniques used to support such questions aim at providing generic and objective methods to process large volumes of data. We claim that the success of a digital humanities project depends on the extent it succeeds in making an appropriate match of the specific with the generic, and the subjective with the objective. Trust in the outcome of a digital humanities study should be informed by a proper understanding of this match, but involves a non-trivial fit for use assessment.

Modern computational tools are often more advanced and exhibit better performance than their less advanced predecessors. These performance improvements come, however, with a price in terms of increased complexity and diminished understanding of what actually happens inside the black box that many tools have become.

Our argument is that scholars in the humanities need to understand the tools they use to the extent that they can assess the limitations of each tool, and how these limitations might impact the outcomes of the study in which the tool is deployed. This level of understanding is necessary, both to be able to assess to what extent the generic task for which the tool has been designed fits the specific problem being studied, and to assess to what extent the limitations of a tool in the context of this specific problem may result in unanticipated side effects that lead to unintended bias or errors in the analysis that threaten the assumed objectivity of each computational step.



```
:- use_rendering(url).
:- set_prolog_stack(global, limit(10 000 000 000)).
:- include(caching).

%%  RECORD PREDICATES

record_date(Nr, Date) :-
    logrecord(Nr, Rec),
    format_time(string(Date), '%+', Rec.time).

record_time(Nr, Time) :-
    logrecord(Nr, Rec),
    Time = Rec.time.

record_url(Nr, URL) :-
    logrecord(Nr, Rec),
    URL = Rec.url().

record_ip(Nr, IP) :-
    logrecord(Nr, Rec),
    IP = Rec.ip.

record_referrer(Nr, Referrer) :-
    logrecord(Nr, R),
    atom_string(Referrer, R.referrer).

record_agent(Nr, Agent) :-
    logrecord(Nr, R),
    Agent = R.agent.

record_status(Nr, Status) :-
    logrecord(Nr, R),
    Status = R.status.

%% generalized for any logrecord with query
%  NOTE: a query can also be empty
%  but still have facets
%  Query_String is of type atom!!
```

```
record_query(Nr, Full_Query, Query_String).
Full_Query = [(query = 'stichting tot uitzending van reclame'), (coll=platform)],
Nr = 14,
Query_String = "stichting tot uitzending van reclame"
```

```
?- record_query(Nr, Full_Query, Query_String).
```

Screenshot of a SWISH executable notebook disclosing the full computational workflow in a search log analysis use case.

Our approach to address this problem is to work closely with research partners in the humanities, and to build together the next generation of e-infrastructures for the digital humanities that allow scholars to make these assessments in their daily research.

This requires more than technical work. For example, to raise awareness for this need, we have recently teamed up with partners from the Utrecht Digital Humanities Lab and the Huygens ING to organise a second workshop in July 2017 around the theme of tool criticism. A recurring issue during these workshops was the need for humanity scholars to be better trained in computational thinking, while at the same time computer scientists (including big data analysts and e-science engineers) need to better understand what information they need to provide to scholars in the humanities to make these fit for use assessments possible.

In computer science, the goal is often to come up with generic solutions by abstracting away from overly specific problem details. Assessing the behaviour of an algorithm in such a specific context is not seen as the responsibility of the computer scientist, while the black box algorithms are insufficiently transparent to transfer this responsibility to the humanities scholar. Making black box algorithms more transparent is thus a key challenge and a sufficiently generic problem to which computer scientists like us can make important contributions. In this context, we collaborate in a number of projects with the Dutch National Library (KB) to measure algorithmic bias and improve algorithmic transparency. For example, in the COMMIT/ project, CWI researcher Myriam Traub is investigating the influence of black box optical character recognition (OCR) and search engine technology on the accessibility of the library's extensive historical newspaper archive [1]. Here, the central question is: can we trust that the (small) part of the archive that users find and view through the online interface is actually a representative view of the entire content of the archive? And, if the answer is no, is this bias caused by a bias in the interest of the users or of a bias implicitly induced by the technology? Traub et al. studied the library's web server logs with six months of usage data, a period in which around a million unique queries had been issued. Traub concluded that indeed only a small fraction (less than 4%) of the overall content of the archive was actually viewed, and that this was not a representative sample at all. The bias however, was largely attributed to a bias in the interest of the users, with a much smaller bias caused by the "preference" of the search engine for medium length articles, which leads to an underrepresentation of overly short and long articles in the search result. Follow-up research is currently focussing on the impact of the OCR on retrievability bias.

In the ERCIM-coordinated H2020 project VRE4EIC [L1] we collaborate with FORTH, CNR and other partners on interoperability across virtual research environments. At CWI, Tessel Bogaard, Jan Wielemaker and Laura Hollink are developing software infrastructure [2] to support improved transparency in the full data science pipeline: involving data selection, cleaning, enrichment, analysis and visualisation. Part of the reproducibility and other aspects related to transparency might be lost if these steps are spread over multiple tools and scripts in a badly documented research environment. Again using the KB search logs as a case study, another part of the challenge lies in creating a transparent workflow on top of a dataset that is inherently privacy sensitive. For the coming years, the goal is to create trustable computational workflows, even if the data necessary to reproduce a study is not available.

Link:
[L1] https://www.vre4eic.eu

References:

[1] M. C. Traub, et al.: "Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus", JCDL 2016: 7-16.
[2] J. Wielemaker, et al.: "Cliopatria: A Swi-Prolog infrastructure for the Semantic Web. Semantic Web", 7(5), 529-541, 2016.

Please contact:

Jacco van Ossenbruggen
CWI, The Netherlands
+31 20 592 4141
Jacco.van.Ossenbruggen@cwi.nl

The DARIAH ERIC: Redefining Research Infrastructure for the Arts and Humanities in the Digital Age

by Jennifer Edmond (Trinity College Dublin), Frank Fischer (National Research University Higher School of Economics, Moscow), Michael Mertens (DARIAH EU) and Laurent Romary (Inria)

As it begins its second decade of development, the Digital Research Infrastructure for the Arts and Humanities (DARIAH) continues to forge an innovative approach to improving support for and the vibrancy of humanities research in Europe.

When we think of infrastructure, we often fall back on canonical images of things like roads, bridges and buildings. In many disciplines, these still resonate with the needs of the researcher community, where a supercollider or a research vessel may indeed be at the core of what is required to advance our state of knowledge.

The arts and humanities are different. As a collection of approaches to knowledge, the methods deployed stem from a shared respect for complex source material emerging not from the experimental design of the scientist, but from the experiences, cultures and creative impulses of human beings. Providing an enhanced, shared, baseline access to key methods, sources, tools and services is therefore a great challenge indeed.

The Digital Research Infrastructure for the Arts and Humanities (DARIAH) was first conceptualised in late 2005 as a response to how this very different set of requirements was being addressed in the fast-moving environment of digitally-enhanced research. The infrastructure was later officially founded as a European Research Infrastructure Consortium (or ERIC) based in France, but with 17 national members contributing funds and in-kind contributions from their local digital humanities research communities. The knowledge base of the resulting network is further enhanced by contributions from funded research projects in which DARIAH is a partner, as well as the contributions of working groups, assembled by network members on a voluntary basis to address key gaps in infrastructural provision or key emerging challenges for the research communities.

This rich tapestry of contributions creates a form of infrastructure based on knowledge production and exchange, rather than on concrete shared facilities or datasets. As such, the challenge for DARIAH as it enters its second decade of development is to capitalise on the human infrastructure it has built to create a fully aligned system of coordinated contribution and access provision to the good practices emerging from the network [1]. In order to do this, we are focussing for the next three years on four key areas of development that will enhance our ability to deliver low-friction, high value interactions for our partners.

The first area of focus is to improve our external communications. Ensuring that our basic information is available in an easily legible form for all current and potential new members of our network is a sine qua non for ensuring that the infrastructure can function. This programme of activities will cover the gamut of communications instruments, from a newly redesigned website to the appointment of specific individual ambassadors to reach key target regions and communities; from a more strategic approach to attending events to a mapping of key organisation-to-organisation relationships within our community and beyond. We will also clarify what we are able to offer as services to our community, from support for grant capture to hosting of orphan research projects. By focussing on our core messages in this way, we hope also to be able to communicate and build consensus around an even clearer message of what DARIAH is and does, and how it operates as an infrastructure in conditions that require a very different approach.

The second plank in DARIAH's development plan is to push forward its vision for a virtual marketplace, making visible and accessible the many tools, services, datasets and expertise bases that our network has opened up for use by others. This may sound like an easy task to achieve, but the prerequisite understanding of what these assets would be valued for and by whom is actually quite challenging to develop. Ensuring that we provide an optimised platform for targeted and serendipitous discovery of resources, as well as their easy reuse, will be a major achievement of DARIAH by 2020.

Our third area of focus is on teaching and learning. Too much focus in the digital humanities is on either training via formal degree programmes or through individual learning via generic platforms like Code Academy or Software Carpentry. The research infrastructure provides a unique environment and set of opportunities for different kinds of learning, aligned to support the individual and institutional modes, but also to provide unique opportunities for experiences of professional acculturation in applied contexts [2]. Already in this area we have active services under the banners of dariahTeach, a Moodle-based, ECTS-linked set of modules, and through our infrastructure cluster project PARTHENOS, which involves the CLARIN ERIC and projects such as IPERION, CHARISMA, ARIADNE, EHRI and CENDARI as well, and where the modules are more targeted at self-learners and as "train-the-trainers" resources. We will now build on these platforms and momentum.

Finally, we view the development of our foresight and policy leadership capacity as a key asset, not only for our current cohort of active digital humanists, but also for the "long tail" of the research communities. These researchers, who may not realise how important the digital is becoming for how they conduct and communicate their research, are a key community for our future growth, and ensuring that we represent their needs at a high level is of central importance to us as we consolidate our position. Our work to date in this area has been linked to initiatives such as DARIAH's contributions to the European Commission's Open Science Policy Platform, and also through our championing of a "Data Reuse Charter" between researchers and cultural heritage institutions, able to promote data sharing and fluidity [3].

On the basis of these interventions, DARIAH is poised to move into its second decade with a reputation as a leader for the arts and humanities, as well as an innovator in research infrastructure.

Link:
[L1] www.dariah.eu

References:

[1] T. Blanke, C. Kristel, L. Romary: "Crowds for Clouds: Recent Trends in Humanities Research Infrastructures" in Cultural Heritage Digital Tools and Infrastructures, A. Benardou, E, Champion, C. Dallas, and L. Hughes eds. Taylor & Francis Group, 2018. https://hal.inria.fr/DARIAH/hal-01248562

[2] G. Rockwell and S. Sinclair: "Acculturation and the Digital Humanities Community", in Digital Humanities Pedagogy: Practices, Principles and Politics, D. Hirsch ed. Cambridge: Open Book, 2012, pp. 177-211.

[3] L. Romary, M. Mertens and A. Baillot, "Data fluidity in DARIAH – pushing the agenda forward," BIBLIOTHEK Forschung und Praxis, De Gruyter, 2016, 39 (3), pp.350-357.

Please contact:

Jennifer Edmond, Trinity College Dublin, Ireland
jedmond36@gmail.com

DIGILAB: A New Infrastructure for Heritage Science

by Luca Pezzati and Achille Felicetti (INO-CNR)

The European Research Infrastructure for Heritage Science, E-RIHS [ˈīris], is working to launch DIGILAB: the new data and service infrastructure for the heritage science research community. First services are expected to be online in 2018.

The European Strategic Roadmap for Research Infrastructures (ESFRI Roadmap [L2]), initiated in 2016, is one of the six new projects of the European Research Infrastructure for Heritage Science (E-RIHS)[1]. E-RIHS supports research on heritage interpretation, preservation, documentation and management. Both cultural and natural heritage are addressed: collections, buildings, archaeological sites, digital and intangible heritage. E-RIHS is a distributed research infrastructure: it includes facilities from many countries, organised in national networks and coordinated by national hubs. The E-RIHS Headquarters – to be seated in Florence, Italy – will provide the unique access point to all E-RIHS services.

E-RIHS will provide state-of-the-art tools and services to support cross-disciplinary research communities of users through its four access platforms (Figure 1):

- MOLAB: offers access to advanced mobile analytical instrumentation for diagnostics of heritage objects, archaeological sites and historical monuments. The MObile LABoratories will allow its users to carry out complex multi-technique diagnostic projects, allowing effective in situ investigation.
- FIXLAB: provides access to large-scale and specific facilities with unique expertise in heritage science, for cutting-edge scientific investigation on samples or whole objects, revealing micro-structures and chemical composition, giving essential and invaluable insights into historical technologies, materials and species, their context, chronologies, alteration and degradation phenomena.
- ARCHLAB: enables physical access to archives and collections of prestigious European museums, galleries, research institutions and universities containing non-digital samples and specimens and organised scientific information.
- DIGILAB: facilitates virtual access to tools and data hubs for heritage research – including measurement results, analytical data and documentation – from large academic as well as research and heritage institutions.

E-RIHS will help the preservation of the world's heritage by enabling cutting-edge research in heritage science, liaising with governments and heritage institutions to promote its continual development and, finally, raising public awareness about cultural and natural heritage and its historic, social and economic significance.

In February 2017, E-RIHS started its preparatory phase supported by the EU project E-RIHS PP (H2020-INFRADEV-02-2016). Representatives of 16 countries (15 from the EU plus Israel) are working together to prepare E-RIHS to be launched as a standalone research infrastructure consortium in 2021.

The DIGILAB platform will provide remote services to the heritage science research community but will also be relevant to and accessible by professionals, practitioners and heritage managers. DIGILAB will enable access to research information as well as to general documentation of analyses, conservation, restoration and any other kind of relevant information about heritage research and background references, such as controlled vocabularies, authority lists and virtual reference collections.

The DIGILAB design takes into account and complies with the EU poli-

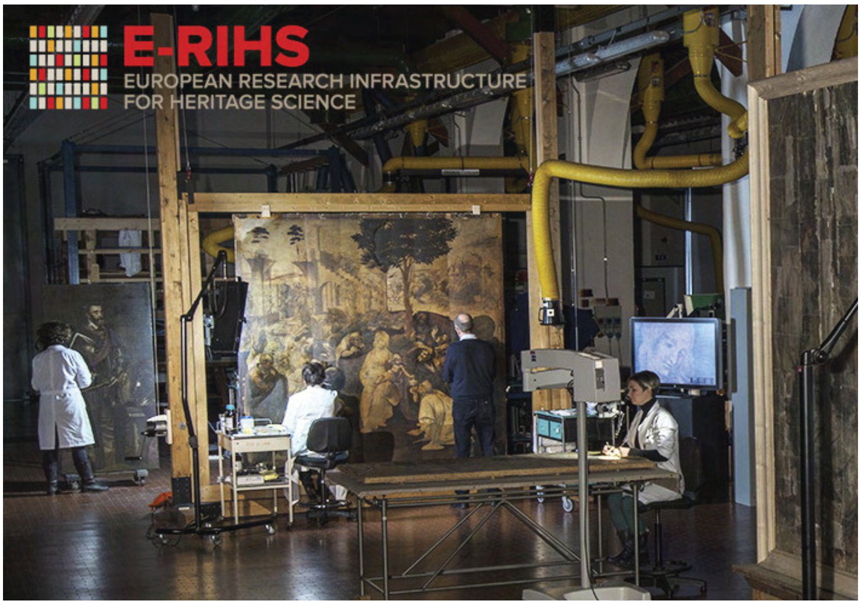
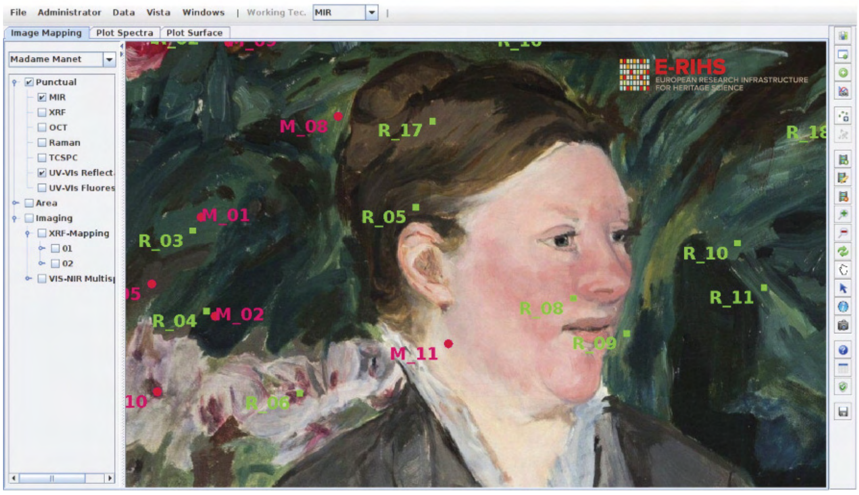


Figure 1: Analytical instrumentation for scientific investigation of heritage objects in E-RIHS.



Working Tec. MIR

Madame Manet

- ☑ Punctual
  - ☑ MIR
  - □ XRF
  - □ OCT
  - □ Raman
  - □ TCSPC
  - ☑ UV-Vis Reflect
  - □ UV-Vis Fluores
- □ Area
- □ Imaging
  - □ XRF-Mapping
    - □ 01
    - □ 02
  - □ VIS-NIR Multis

Figure 2: A DIGILAB service for mapping and comparing analysis results to be used for painting restoration and preservation.

cies and strategies concerning scientific data, including the FAIR [L3] data principles [1], the Open Research Data [L4] policy, and the EOSC [L5] strategy. DIGILAB will rely on a network of federated repositories where researchers, professionals, managers and other heritage-related professionals deposit the digital results of their work. DIGILAB will not keep those data internally: instead, it will provide access to the original repositories where the data are stored.

DIGILAB is inspired by the FAIR data principles: it will enable Finding data through an advanced search system operating on a registry containing metadata describing each individual dataset; it will support Access to the data through a federated identity system, while data access grants will be local to each repository; it will guarantee data Interoperability by requiring the use of a standard data model; finally, it will foster Re-use by making services available to users, to process the data according to their own research questions and use requirements.

Data policies are relatively new to the heritage science sector. In many cases digital data are still considered disposable: they are used as long as the research continues, then disposed of. They are often stored in inaccessible systems, like personal hard disks or other storage devices, and are thus quickly rendered unusable through the lack of proper documentation, the obsolescence of the software used to create and manage them or, simply, the degradation of storage hardware, which after some time becomes unreadable either because it ceases to function or because it can no longer be connected to more advanced devices.

On the other hand, most of this valuable information can be easily recovered and organised into usable archives. This often requires a handicraft approach, tailored to each dataset and dependent on the humans who created it and who are still available to provide the necessary information. DIGILAB will set up guidelines for dataset recovery and assist researchers and research institutions willing to undertake such tasks.

The cloud-based DIGILAB infrastructure, enhanced by the adoption of international standards and modern IT solutions, will provide the necessary flexibility for the efficient aggregation, interoperability implementation and integration management of huge amounts of scientific data, in order to foster their publication and redistribution in various formats, in accordance with the related policies. The complex semantic graph built in the DIGILAB registry will be able to trace new research paths through scientific concepts by means of the efficient use of semantic relationships linking the entities involved in scientific research; this will provide DIGILAB users with new tools for the discovery, access, and re-use of relevant information (Figure 2).

DIGILAB is designed to be the privileged gateway to European scientific knowledge in heritage, in preparation for becoming the main international portal for heritage science research.

Links:

[L1] http://www.e-rihs.eu/
[L2] http://www.esfri.eu/
[L3] https://kwz.me/hTO
[L4] https://kwz.me/hm0
[L5] https://kwz.me/hm1

Reference:

[1] M. D. Wilkinson, et al.: “The FAIR Guiding Principles for scientific data management and stewardship”, Sci. Data 3:160018, 2016, doi: 10.1038/sdata.2016.18

Please contact:

Luca Pezzati, Achille Felicetti, CNR INO – Istituto Nazionale di Ottica, Italy, +390552308279,
luca.pezzati@cnr.it

Building a Federation of Digital Humanities Infrastructures

by Alessia Bardi and Luca Frosini (ISTI-CNR)

Research infrastructures (RIs) are "facilities, resources and services used by the science community to conduct research and foster innovation" [1]. Researchers' needs for digital services led to the realisation of e-Infrastructures, i.e., RIs offering digital technologies for data management, computing and networking. Relevant examples are high speed connectivity infrastructures (e.g., GÈANT), grid computing infrastructures (e.g., European Grid Infrastructure EGI), scholarly communication infrastructures (e.g., OpenAIRE), data e-infrastructures (e.g., D4Science).

Digital humanities infrastructures (DHIs) are e-infrastructures supporting researchers in the field of humanities with a digital environment where they can find and use ICT tools and research data for conducting their research activities. A growing number of DHIs have been realised, most of them targeting a specific sector of humanities, such as ARIADNE [L2] for archeology, EHRI [L3] for studies on the holocaust, Cendari [L4] for history, CLARIN [L5] for linguistic, and DARIAH [L6] for arts and humanities. Thanks to their discipline-specific feature, those DHIs offer specialised services and tools to researchers, who are now demanding support for interdisciplinary research, common solutions for data management, and access to resources that are traditionally relevant to different sectors (e.g., text-mining algorithms traditionally used by linguists can also be useful to historians and social scientists).

One of the main goals of the PARTHENOS project (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies – EC-H2020-RIA grant agreement 654119) is to bridge existing DHIs by forming a federation where researchers of different sectors of the humanities can collaborate and share data, services and tools in an integrated environment.

PARTHENOS will produce a complete technical framework for the federation, enabling transparent access to resources managed by different DHIs and enabling the creation and operation of virtual research environments [1] where researchers with different backgrounds can collaborate on specific research topics.

The technical framework supports the realisation of the federation by offering tools for:

- The creation of an homogenous information space where all resources (data, services and tools) of the different DHIs are described according to a common data model.
- The discovery of available resources.
- The use of available resources (for download or processing).
- The creation of VREs where users can find resources relevant for a research topic, run services, and share the computational results.

The technical framework (see Figure 1) includes two main components: the PARTHENOS Content Cloud Framework (CCF) and the Joint Resource Registry (JRR).

The CCF supports the aggregation of metadata about resources from the DHIs of the federation. The PARTHENOS aggregator is realised with the D-NET software toolkit [2], an enabling framework for the realisation of Aggregative Data Infrastructures (ADIs) developed and maintained by CNR-ISTI. D-NET provides functionality for the automatic collection, harmonisation, curation and delivery of metadata coming from a dynamic set of heterogenous data providers. In the context of the PARTHENOS project, D-NET has been configured to collect metadata made available by existing DHIs operated by PARTHENOS partners (namely: ARIADNE, CENDARI, CLARIN, CulturaItalia, DARIAH DE, DARIAH GR/DYAS, DARIAH IT, EHRI, Huma-Num, ILC) and harmonise them according to an extension of the CIDOC-CRM model [L7] [L8] by applying X3ML [L9] mappings. The mapping language, editor and execution

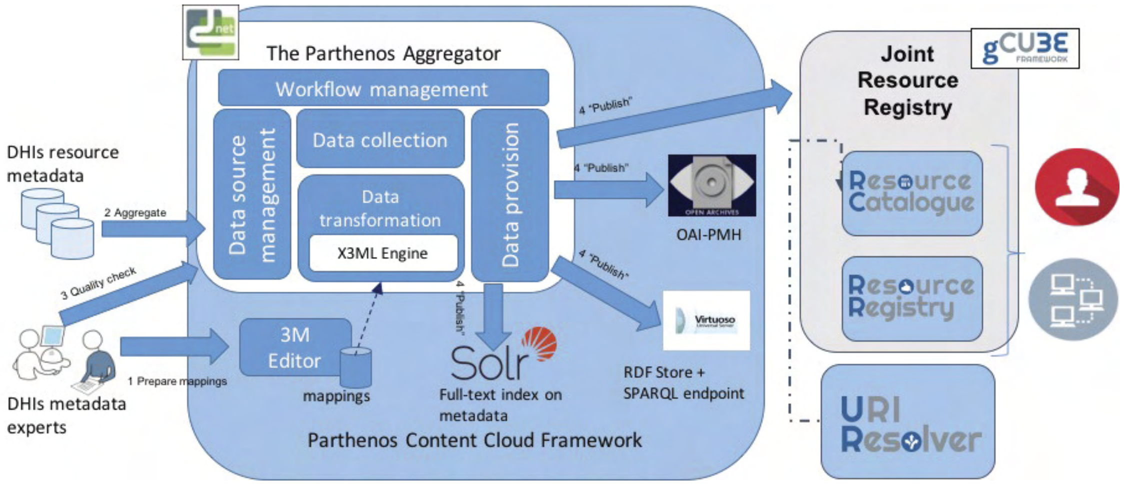


Figure 1: Technical framework for DHIs federation.

engine are realised and maintained by the Greek partner FORTH. Aggregated content is then published via different endpoints, supporting a set of (de-facto) standard protocols for metadata search (Solr API, SPARQL) and exchange (OAI-PMH).

The aggregated content is also ingested into the Joint Resource Registry, which exposes an end-user GUI (Resource Catalogue) and a machine-oriented API (Resource Registry) for resource discovery. Data and services registered in the JRR become discoverable by and accessible to users and other services of the federation. Moreover JRR provides functionality for infrastructure management. For example, a user can run a CLARIN service for full-text mining on a dataset of medieval full-texts available in the CENDARI DHI in a transparent way. Computational results can be easily stored and shared with a selection of colleagues or publicly, by publishing them into the JRR.

The JRR is based on the gCube enabling technology [3], an open-source software toolkit used for building and operating hybrid data infrastructures [4] enabling the dynamic deployment of virtual research environments by favouring the realisation of reuse oriented policies. gCube is developed and maintained by CNR-ISTI.

The Parthenos technical framework is currently at the beta stage and operated on the D4Science infrastructure [L10] at the Institute of Information Science and Technologies of the Italian National Research Council (ISTI-CNR). Representatives of the consortium are actively preparing mappings for metadata, selecting data and services to share and setting up VREs. As of August 2017, two VREs have been created: one includes services for natural language processing and semantic enrichment of textual data; the other is meant for the integration of reference resources. In the coming months, the framework will be deployed in a production environment and assessed by a selection of humanities researchers in the consortium. We plan to open the framework to all researchers of DHIs in the consortium by the end of the project (April 2019).

References:

[1] L. Candela et al.: "Virtual research environments: an overview and a research agenda", Data Science Journal, 12, GRDI75-GRDI81, 2013. http://doi.org/10.2481/dsj.GRDI-013

[2] P. Manghi et al. "The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures", Program, Vol. 48 Issue: 4, pp.322-354, 2014 doi: https://doi.org/10.1108/PROG-08-2013-0045

[3] L. Candela, P. Pagano: "Cross-disciplinary data sharing and reuse via gCube", in: ERCIM News, Issue 100, January 2015. https://kwz.me/hO7

[4] L. Candela et al.: "Managing Big Data through Hybrid Data Infrastructures", in ERCIM News, Issue 89, April 2012. https://kwz.me/hO8

Links:

[1] https://kwz.me/hO9
[2] http://www.ariadne-infrastructure.eu/
[3] https://www.ehri-project.eu/
[4] http://www.cendari.eu
[5] https://www.clarin.eu
[6] http://www.dariah.eu/
[7] http://www.cidoc-crm.org/
[8] https://kwz.me/hOf
[9] https://kwz.me/hOj
[10] https://parthenos.d4science.org/

Please contact:

Alessia Bardi, Luca Frosini
ISTI-CNR, Italy
alessia.bardi@isti.cnr.it,
luca.frosini@isti.cnr.it

Knowledge Complexity and the Digital Humanities: Introducing the KPLEX Project

by Jennifer Edmond and Georgina Nugent Folan (Trinity College Dublin)

The KPLEX project is looking at big data from a rich data perspective. It uses humanities knowledge to explore bias in big data approaches to knowledge creation.

The KPLEX Project is an H2020 funded project tasked with investigating the complexities of humanities and cultural data, and the implications of digitisation on the unique and complex messy data that humanities and cultural researchers are accustomed to dealing with. The drive for ever greater integration of digital humanities (DH) data is complicated by the uncomfortable truth that a lot of the information that should be the cornerstones of our decision making, rich data about the history of our economies, societies and cultures, isn't digitally available. This, along with the "epistemics of the algorithm"[1] are key concerns of the KPLEX project, and we are working to expand awareness of the risks inherent in big data for DH and cultural research, and to suggest ways in which phenomena that resist datafication can still be represented (if only by their absence) in knowledge creation approaches reliant upon the interrogation of large data corpora.

KPLEX addresses the repercussions of the dissociation of data sources from the people, institutions and conditions that created them. In a rapidly evolving DH environment where large scale data aggregation is becoming ever more accepted as the gold standard, the K-PLEX project is defining and describing some of the key aspects of data that are at risk of being left out of our knowledge creation processes, and the strategies researchers have developed to deal with these complexities.[2]

The K-PLEX team is diverse and has adopted a comparative, multidisciplinary, and multi-sectoral approach to his problem, focussing on four key challenges to the knowledge creation capacity of big data approaches:

1) redefining what data is and the terms we use to speak about it [3];

2) the manner in which data that are not digitised or shared become “hidden” from aggregation systems;
3) the fact that data is human created, and lacks the objectivity often ascribed to the term;
4) the subtle ways in which data that are complex almost always become simplified before they can be aggregated.

We approach these questions with a humanities research perspective and remain committed to humanities methodologies and forms of knowledge, but we make use of social science research tools to look at both the humanistic and computer science approaches to the term “data” and its many possible meanings and implications. Our core shared discourse of the digital humanities allows us to use these methods and knowledge in a contextualised, socially relevant manner, a strength of our consortium that is further enhanced by our inclusion of both ethnographic/anthropological and industrial perspectives.

Led by Trinity College Dublin, the KPLEX team spans four countries, taking in Freie Universität Berlin (Germany), DANS-KNAW (The Hague) and TILDE (Latvia). Each of the K-PLEX project partners addresses an integrated set of research questions and challenges. The research teams have been assembled to pursue a set of questions that are humanist-led, but broadly interdisciplinary, including humanities and digital humanities, data management, anthropology and computer science, but also including stakeholders from outside of academic research able to inform the project’s evidence gathering and analysis of the challenges, including participation from both a technology SME (TILDE) and a major national ICT research centre (ADAPT, Ireland). In addition, KPLEX takes in the experiences of a large number of major European digital research infrastructure projects federating cultural heritage data for use by researchers, through the contributions by TCD (Dublin) and KNAW-DANS (The Hague). These projects (including CENDARI, EHRI, DARIAH-EU, DASISH, PARTHENOS, ARIADNE and HaS) have all faced and progressed the issues surrounding the federation and sharing of cultural heritage data. In addition, two further projects that deal with non-scientific aspects of researcher epistemics are also engaged, namely the “Scholarly Primitives and Renewed Knowledge Led Exchanges” project (SPARKLE, based at TCD) and the “Affekte der Forscher” (based at FUB). These give the KPLEX team and project a firm baseline of knowledge for dealing with the question of how epistemics creates and marks data.

The KPLEX project kicked off in January 2017, and will conclude in March 2018, presenting its results via a composite white paper that unites the findings of each research team, with each research team also producing a peer reviewed academic paper on their findings. Over the coming months the project will be represented at DH conferences in Liverpool (“Ways of Being in the Digital Age”), Austria (“Data First!? Austrian DH Conference”), Manchester (“Researching Digital Cultural Heritage International Conference”) and Tallin (“Metadata and Semantics Research Conference”).

Links:
[L1] https://kplex-project.com/,
Twitter: @KPLEXProject,
Facebook: KPLEXProject

References:

[1] T Presner: “The Ethics of the Algorithm”, in Probing the Ethics of Holocaust Culture.
[2] J Edmond: “Will Historians Ever Have Big Data?” In Computational History and Data-Driven Humanities. doi:10.1007/978-3-319-46224-0_9.
[3] L Gitelman, ed., “‘Raw Data’ is an Oxymoron”.

Please contact:

Jennifer Edmond, Georgina Nugent Folan, Trinity College Dublin, Ireland
edmondj@tcd.ie, nugentfg@tcd.ie

Restoration of Ancient Documents Using Sparse Image Representation

by Muhammad Hanif and Anna Tonazzini (ISTI-CNR)

Archival, ancient manuscripts constitute a primary carrier of information about our history and civilisation process. In the recent past they have been the object of intensive digitisation campaigns, aimed at their preservation, accessibility and analysis. At ISTI-CNR, the availability of the diverse information contained in the multispectral, multisensory and multiview digital acquisitions of these documents has been exploited to develop several dedicated image processing algorithms. The aim of these algorithms is to enhance the quality and reveal the obscured contents of the manuscripts, while preserving their best original appearance according to the concept of “virtual restoration”. Following this research line, within an ERCIM “Alain Bensoussan” Fellowship, we are now studying sparse image representation and dictionary learning methods to restore the natural appearance of ancient manuscripts affected by spurious patterns due to various ageing degradations.

The collection of ancient manuscripts serves as history’s own closet, carrying stories of enigmatic, unknown places or incredible events that took place in the distant past, many of which are yet to be revealed. These manuscripts are of great interest and importance for historians to study people of the past, their culture, civilisation and way of life. Most of the ancient classic documents have had a very narrow escape from total annihilation. Thus, digital preservation of our documental heritage has been one of the first focusses of the massive archive and library digitisation campaigns performed in the recent years. This, in turn, has contributed to the birth of digital humanities as a science. In addition to preservation, computing technologies applied to the digital images of these documents have quickly become a powerful and versatile tool to simplify their study and retrieval, and to facilitate new insights into the documents' contents.

The quality of the digital records, however, depends on the current status of the original manuscripts, which in most cases are affected by several types of degradation, such as spots, ink fading, or ink seeping from the reverse side, due to bad storage conditions and the fragile nature of the materials (e.g., humidity, mould, ink chemical properties, paper porosity). In particular, the phenomenon of ink seeping is perhaps the most frequent one in ancient manuscripts written on both sides of the sheet. This effect, termed as bleed-through, becomes visible as an unpleasant and disturbing degradation pattern, severely impairing legibility, aesthetics, and interpretation of the source document.

In general, bleed-through removal is addressed as a classification problem, where image pixels are labelled as either background (paper texture), bleed-through (seeped ink), or foreground (original text). This classification problem is very difficult, since the intensity of both foreground text and bleed-through pattern can be so highly variable as to make it extremely hard or even impossible to distinguish them. In addition, when the aim is to obtain a very accurate and plausible restoration, another very critical issue is to replace the identified bleed-through pixels with appropriate replacement colour values, which do not alter the original look of the manuscript.

Recently, we proposed a two-step method to address bleed-through document restoration from a pre-registered pair of recto and verso images of the manuscript. First, the bleed-through pixels are identified on both sides [1]; then, a sparse representation based image inpainting technique is applied to fill-in the bleed-through pixels, by taking into account their propensity to aggregation, and in accordance with the natural texture of the surrounding background.

Sparse representation methods are reported with state-of-the-art results in different image processing applications. These methods process the whole image by operating on a patch-by-patch level. For this specific application, our aim is to reproduce the background texture to maintain the original look of the document. In the sparse representation setup, an over-complete dictionary is learned using a set of training patches from the recto and verso image pair. In the training set we only select patches with no bleed-though pixels. This choice speeds up the training process since it excludes non-informative image regions. For each patch to be inpainted, we first search for its mutual similar patches in a small bounded neighbourhood window. We used a block matching technique with Euclidean distance metric as similarity criterion. The similar patches are grouped together in a matrix. A group-based sparse representation method [2] is exploited to find the befitting fill-in for the bleed-through strokes. The use of similar patch groups incorporates local information that helps to preserve the natural colour/texture continuation property of the physical manuscript.

An original degraded manuscript and its restored version are presented in Figure 1. It is worth noting that our algorithm can be directly applied to inpaint any other possible interference pattern detected in the paper support (e.g., stains). We are also studying the extension of the method to the restoration of broken or faded foreground characters.

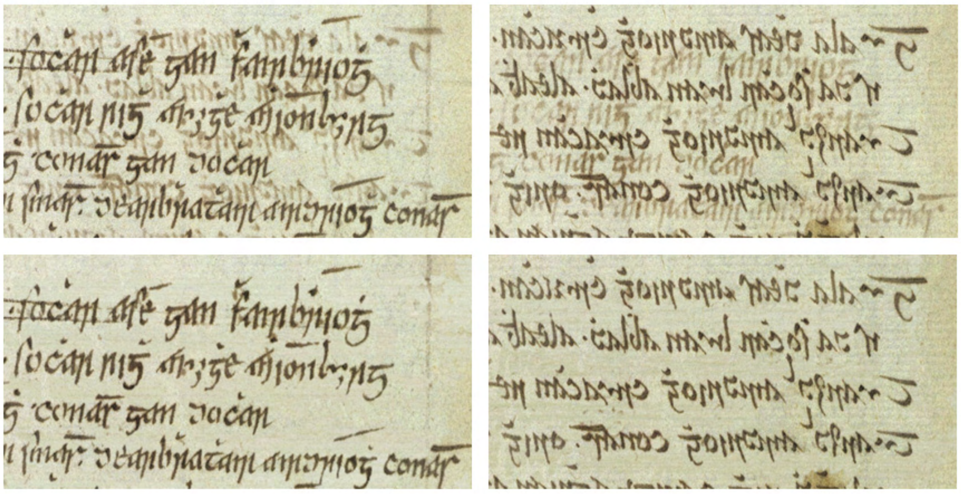


Figure 1: A visual comparison of an original ancient manuscript effected by bleed-through degradation and its restored version using our method. The first row shows the degraded recto and verso pair, and the restored images are presented in the second row.

References:

[1] A. Tonazzini, P. Savino, and E. Salerno: "A nonstationary density model to separate overlapped texts in degraded documents," Signal, Image and Video Processing, vol. 9, pp. 155–164, 2015.

[2] J. Zhang and D. Zhaocand W. Gao: "Group-based sparse representation for image restoration," IEEE Trans. Image Process., vol. 32, pp. 1307–1314, 2016.

Please contact:

Anna Tonazzini, ISTI-CNR, Pisa
+39 3483972150
anna.tonazzini@isti.cnr.it

St Paul's Cathedral Rises From the Dust – News From the Virtual St Paul's Cathedral Project

by John N. Wall (NC State University), John Schofield (St Paul's Cathedral, London), David Hill (NC State University and Yun Jing (NC State University)

The Virtual St Paul's Cathedral Project [L1], now at the half-way point in its development, is beginning to show results. Attached are images of our draft model of St Paul's Cathedral and buildings in the cathedral's churchyard, from the early 1620's, before everything seen here was destroyed by the Great Fire of London in 1666. These images are based on a combination of contemporary images of the cathedral and its surrounding buildings, surveys of these buildings made after the Great Fire, and images of appropriate buildings from this period that survive in modern-day England.

When we are done, the visual model will also incorporate details of weather and climate, as well as recent scholarship into the material and social history of London as it grew in the 16th and 17th centuries into a city of over 200,000 people.

Supported by a Digital Humanities Implementation Grant from the National Endowment for the Humanities, the goal of the Virtual St Paul's Cathedral Project is to recreate the experience of worship and preaching in St Paul's Cathedral in London [1] in the early seventeenth century. This project demonstrates the value of visual and acoustic modelling in helping us understand the look and feel of historic sites as well as their acoustic properties, by recreating events that took place in those spaces, to experience these events as they unfold, minute by minute, in the spaces in which they originally took place.

Along the way we are developing an open source acoustic modelling software package to enable others to explore at minimal cost the acoustic properties of other historic sites. Our use of digital technology for the St Paul's Cathedral Project aspires to be scrupulously accurate, integrating into a single experiential model the rich record of information available about St Paul's and its worship.

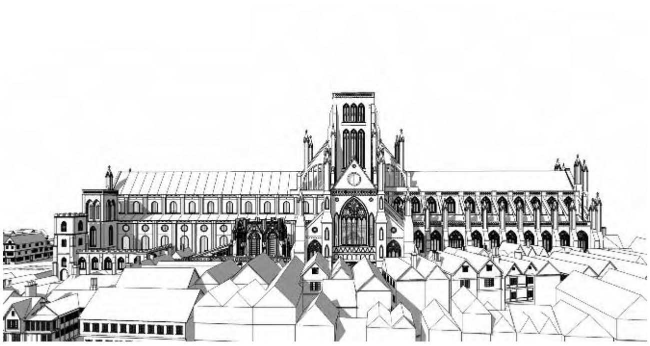


Figure 1: St Paul's Cathedral as seen from the south.

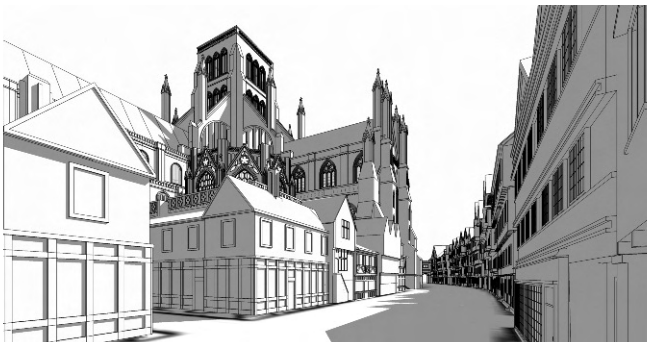


Figure 2: St Paul's Cathedral as seen from the Southwest.

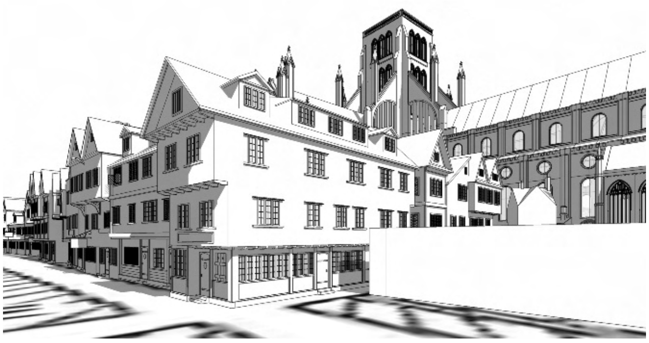


Figure 3: St Paul's Cathedral as seen from the Southeast.

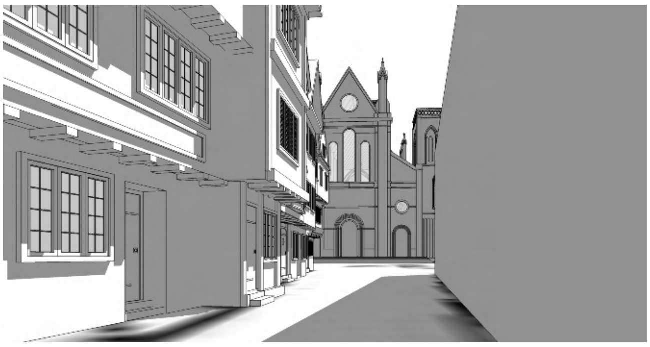


Figure 4: St Paul's Cathedral, the West Front.

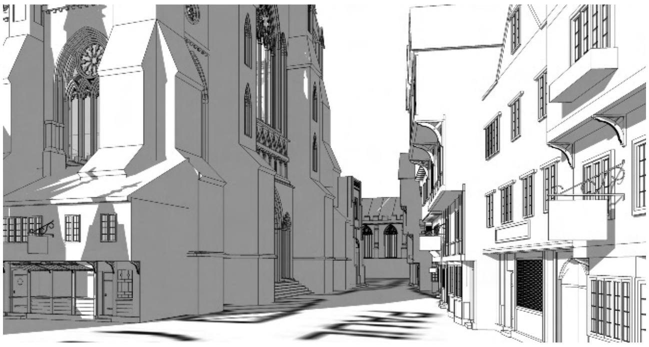


Figure 5: St Paul's Cathedral, the West Front as seen from the Northwest.

We will restage worship services conducted inside and outside this virtual cathedral, drawing on the rites of the Book of Common Prayer (1604) and incorporating music for choir and organ composed for performance in these spaces by musicians at St Paul's in the late 16th and early 17th centuries [2]. The organ pieces will be played on a digital reconstruction of a 17th century instrument that closely approximates the specifications of the organ at St Paul's; the singers and the actors will perform choral pieces, prayers, Bible readings, and the texts of sermons in a linguistic reconstruction of the sound of early modern London speech .

In this virtual environment, we will be able to experience the worship of the post-Reformation Church of England as an unfolding experience in the specific architectural setting and of St Paul's Cathedral and in the context of a responsive and collaborative congregation as well as a noisy crowd, just over the choir screen, of shoppers, merchants, and fashionable people out for gossip and public recognition in Paul's Walk.

This project builds on the successful completion, in 2013, of our proof-of-concept project, the Virtual Paul's Cross Project, funded by a Digital Humanities Level II Start-Up Grant in 2011. The Virtual Paul's Cross Project [3], accessible through its website (http://vpcp.chass.ncsu.edu), makes it possible for us to experience John Donne's sermon for November 5, 1622, performed in its original pronunciation, from eight different listening positions within a virtual model of the historic space of Paul's Churchyard in London and in the presence of four different sizes of crowd.

The Virtual Paul's Cross Project website has been widely recognised for its contribution to research into early modern religious life and has earned two international awards, the John Donne Society's Award for Distinguished Digital Publication (2013) and the Award for Best DH Data Visualization from DH Awards (2014).

Link:
[L1] http://vpcp.chass.ncsu.edu

References:
[1] D. Keene, et al. (ed.): "St Paul's: The Cathedral Church of London, 604-2004, 2004; and J. Schofield: "St Paul's Cathedral Before Wren", English Heritage, 2011.
[2] Sources for choral music include scores now in the archives of the Royal Academy of Music (partially published in 1641 as The First Book of Selected Church Musick, ed. John Barnard, a Minor Canon and member of the choral foundation at St Paul's) and organ music (the so-called "Batten Organ Book) in the Bodleian Library at Oxford.
[3] The Virtual Paul's Cross Project has been widely reviewed, including Matthew J. Smith, "Meeting John Donne: The Virtual Paul's Cross Project," Spenser Review 44 (Fall 2014), at https://kwz.me/hkR

Please contact
John N. Wall,
NC State University, USA
+1 919 515 4162, jnwall@ncsu.edu

Immersive Point Cloud Manipulation for Cultural Heritage Documentation

by Jean-Baptiste Barreau (CNRS/CReAAH UMR 6566), Ronan Gaugne (Université de Rennes 1/IRISA-Inria) and Valérie Gouranton (INSA Rennes/ IRISA-Inria)

A point cloud is the basic raw data obtained when digitizing cultural heritage sites or monuments with laser scans or photogrammetry. These data represent a rich and faithful record provided that they have adequate tools to exploit them. Their current analyses and visualizations on PC require software skills and can create ambiguities regarding the architectural dimensions. We propose a toolbox to explore and manipulate such data in an immersive environment, and to dynamically generate 2D cutting planes usable for CH documentation and reporting.

Due to their states, complexity and archaeological interest, the two subjects studied in this work are Breton architectural sites for which the development of new analytical techniques appears quite appropriate. The first is the chapel of Languidou, built in the middle of the 13th century in the municipality of Plovan, which seems to be the "founding element" of a religious architectural style. The second building addressed in this work is the "jeu de paume" court of Rennes, built at the beginning of the 17th century and registered as a Historical Monument. For several decades, the study of these kinds of architectural styles and reorganizations has been carried out by archaeologists by producing different types of 2D documentation: plans, drawings, sections, profiles and orthophotos. 3D documentation comes from classical topographic survey, laser scanning or photogrammetry. In the case of the French Grand-Ouest, some of this documentation has been carried out within the scope of the West Digital Conservatory of Archaeological Heritage [1]. Until now, an engineer has performed the segmentations of the 3D documentation on a PC and tries to better meet the archaeologist's expectations, who is sometimes unwilling to use new technologies and frustrated by his lack of autonomy. The objective of this work is to involve the archaeologist more deeply in this proces by immersing and allowing him to segment, in real time and on 1:1 scale, the 3D survey.

The first step consists in loading the point clouds of the architectural sites into the Virtual Reality platform Immersia [2]. The scans were done in June 2013 and June 2014, with a Leica ScanStation C10 and a Focus3D X330, and were integrated a few months later (see Figure 1). The data structure of the

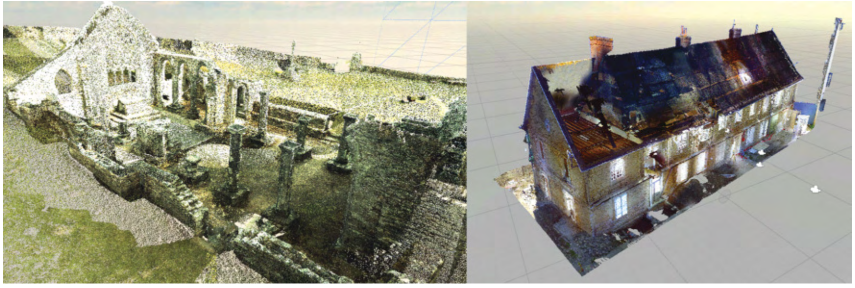


Figure 1: Point clouds of the chapel of Languidou and the "jeu de paume" court.

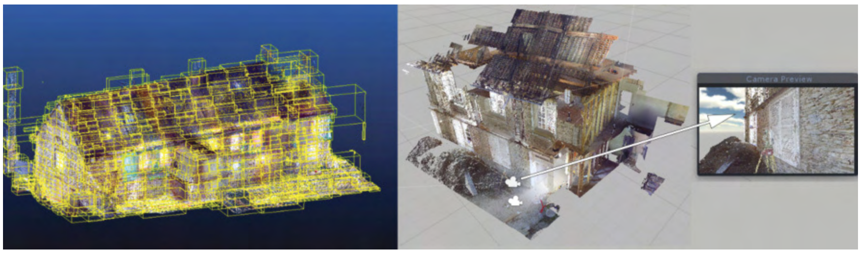


Figure 2: Octree partitioning and culling used on "jeu de paume" court points cloud.

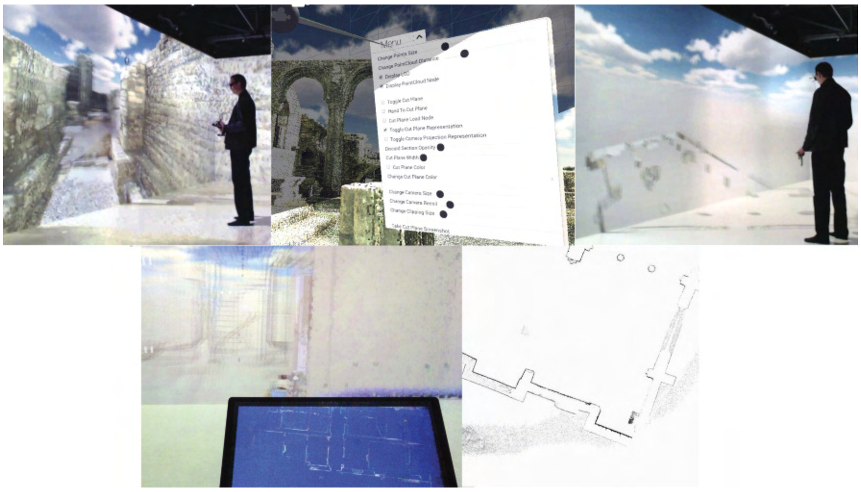


Figure 3: "Cutting plane" creation process by an operator in the Immersia platform.

points displayed in our virtual reality device is a billboard with 3 coordinates, a colour and a scalar. The files are in a binary format that we designed for Unity. For a correct exploration within the Immersia, a subsampling was systematically done and two cloud loading modes were implemented. For the first mode, the cloud is distributed over several hundred octrees. They are loaded dynamically and are visible from the user's point of view, thanks to a culling technique (cf Figure 2). The second mode consists in the use of a Level-of-Detail technique. The number of points loaded at the start in a single file is smaller and their size is fixed. The selection of cloud segmentation tools is done through a MiddleVR menu that allows the user to interact with three parameters (cf Figure 3). When the first mode is active, it is possible to modify the size of the points. On a more global scale, the user can switch between the two loading modes and change the display distance to the cloud. Concerning the "cutting plane", it is possible to display or hide it, change its thickness and assign a unique color to the points contained in it. It is also possible to change the opacity of points outside the cutting plane. Its manipulation within Immersia is done thanks to a Flystick which modifies its translations and rotations. The display in the Immersia platform (MiddleVR) has a frame rate which can reach 30fps. This result is two to three times less fluent than that on PC (Unity / MiddleVR). The rendering of the 2D resulting plane is done with a camera orthogonal to the 3D cutting plane and can displayed on a tablet (cf Figure 3). The user can also adjust the distance, field of view and roll of the camera. A scale is also generated automatically on the resulting 2D plane, in order to provide a correct understanding of the dimensions.

We now need to have the tool tested by a large community of archaeologists, in particular to check if the use of the Flystick to move the cutting plane is sufficiently precise. At the same time, we still have to improve the optimization of the point clouds management. Concerning the reduction of the I/O time and disk space, it will be necessary to store the points coordinates on 2 bits (instead of 4), to downgrade the precision to the millimetre unit, to decrease the entropy, and to use zstd compression. The use of culling occlusion, at native Unity or hardware level, is also being studied to improve optimization. Finally, the management of additional data associated with points is a major issue. Their storage, visualization and linking seems indeed to be a very interesting long-term prospect.

References :

[1] J.-B. Barreau, R. Gaugne, Y. Bernard, G. Le Cloirec, and V. Gouranton, "The West Digital Conservatory of Archaeological Heritage Project," in DH, (France), pp. 547–554, 2013.

[2] R. Gaugne, V. Gouranton, G. Dumont, A. Chauffaut, and B. Arnaldi, "Immersia, an open immersive infrastructure: doing archaeology in virtual reality," Archeologia e Calcolatori, supplemento 5, pp. 180-189, 2014.

Please contact:

Jean-Baptiste Barreau
CNRS/CReAAH UMR 6566, Rennes, France
jean-baptiste.barreau@univ-rennes1.fr

Ronan Gaugne
Université de Rennes 1/IRISA UMR 6074, Rennes, France
ronan.gaugne@irisa.fr

Valérie Gouranton
INSA Rennes/IRISA UMR 6074, Rennes, France
valerie.gouranton@irisa.fr

Culture 3D Cloud: A Cloud Computing Platform for 3D Scanning, Documentation, Preservation and Dissemination of Cultural Heritage

by Pierre Alliez (Inria), François Forge (Reciproque), Livio de Luca (CNRS MAP), Marc Pierrot-Deseilligny (IGN) and Marius Preda (Telecom SudParis)

One of the limitations of the 3D digitisation process is that it typically requires highly specialised skills and yields heterogeneous results depending on proprietary software solutions and trial-and-error practices. The main objective of Culture 3D Cloud [L1], a collaborative project funded within the framework of the French "Investissements d'Avenir" programme, is to overcome this limitation, providing the cultural community with a novel image-based modelling service for 3D digitisation of cultural artefacts. This will be achieved by leveraging the widespread expert knowledge of digital photography in the cultural arena to enable cultural heritage practitioners to perform routine 3D digitisation via photo-modelling. Cloud computing was chosen for its capability to offer high computing resources at reasonable cost, scalable storage via continuously growing virtual containers, multi-support diffusion via remote rendering and efficient deployment of releases.

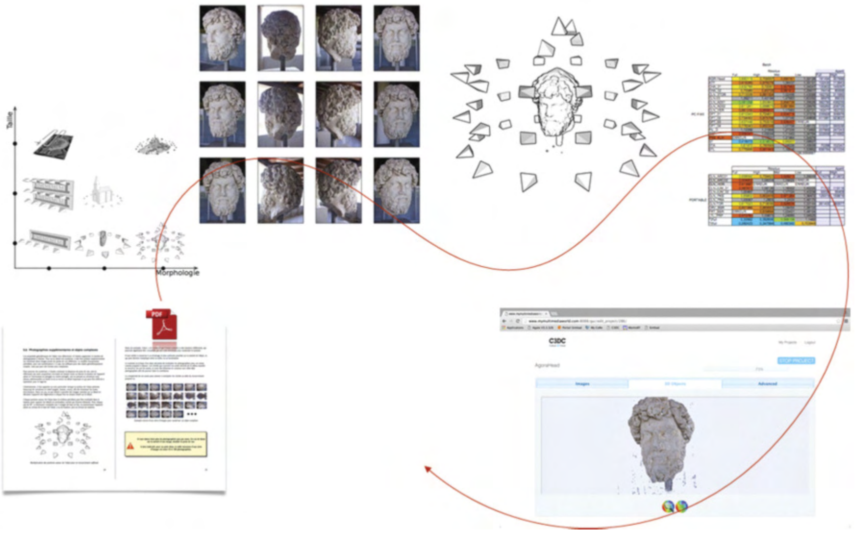
Platform

The platform is designed to be versatile in terms of scale and typologies of artefacts to be digitised, scalable in terms of storage, sharing and visualisation, and able to generate high-definition and accurate 3D models as output. The platform has been implemented from modular open-source software solutions [L2, L3 1, 2, 3]. The current cloud-based platform hosted by TGIR HumaNum [L4] enables four simultaneous users in the form of affordable high performance computing services (8 cores, 64GB of RAM and 80 GB of disk).

Pipeline

The photo-modelling web service offered by the platform implements a modular pipeline with various options depending on the user mode that ranges from novice to expert (Figure 1). The minimum requirement is to conform to several requirements during acquisition: A protocol specific to each type of artefact (statue, façade, building, interior of building) and sufficient overlap between the photos. Image settings such as EXIF, exposure and white-balance can be adjusted or read from the image metadata. The image sequence can be organised into a linear, circular or random sequence. Camera calibration is performed automatically and a dense photogrammetry approach based on image matching is performed to generate a dense 3D point set with colour attributes. For the artefact shown in Figure 1 the Micmac software solution [2] generates a 14M point set from 26 photos. The output point set can be cleaned up from outliers and simplified, and a Delaunay-based surface reconstruction method turns it into a dense surface triangle mesh with colour attributes.

We plan to improve the platform in order to deal with series of multifocal photos, fisheye devices, and photos acquired by unmanned aerial vehicles (drones).



Links:
[L1] http://c3dc.fr/
[L2] http://logiciels.ign.fr/?Micmac
[L3] https://www.cgal.org/ (see components "Point set processing" and "Surface reconstruction"
[L4] http://www.huma-num.fr/

References:

[1] E. RupnikEmail, M. Daakir and M. Pierrot-Deseilligny: "MicMac – a free, open-source solution for photogrammetry", Open Geospatial Data, Software and Standards 2017.

[2] T. van Lankveld: "Scale-Space Surface Reconstruction", in CGAL User and Reference Manual. CGAL Editorial Board, 4.10.1 edition, 2017.

[3] P. Alliez, et al: "Point Set Processing", in CGAL User and Reference Manual. CGAL Editorial Board, 4.10.1 edition, 2017.

Please contact:
Pierre Alliez, Inria, France
pierre.alliez@inria.fr

Livio de Luca, CNRS, France
livio.deluca@map.cnrs.fr

Figure 1:Figure 1: The photo-modelling process requires implementing a well-documented acquisition protocol specific to each type of data, taking a series of high-definition photos, specifying the type of acquisition (linear, circular, random), performing camera calibration (locations and orientations), generating a dense 3D point set with colour attributes via dense image matching and reconstructing a surface triangle mesh.

Physical Digital Access Inside Archaeological Material

by Théophane Nicolas (Inrap/UMR 8215 Trajectoires), Ronan Gaugne (Université de Rennes 1/IRISA-Inria) and Valérie Gouranton (INSA de Rennes/ IRISA-Inria) and Jean-Baptiste Barreau (CNRS/CReAAH UMR 6566)

Traditionally, accessing the interior of an artefact or an archaeological material is a destructive activity. We propose an alternative non-destructive technique, based on a combination of medical imaging and advanced transparent 3D printing.

Our project proposes combining a computed tomography (CT) scan and advanced 3D printing to generate a physical representation of an archaeological artefact or material. This project is conducted in Rennes, France, with archaeologists from Inrap [L1] and computer scientists from Inria [L2]. The goal of the project is to propose innovative practices, methods and tools for archaeology based on 3D digital techniques.

Archaeologists and curators regularly experience the problem of needing to work on objects that are themselves or have features which are inaccessible. For example, artefacts may be encased in corroded materials or in a cremation burial, or integrated in, and inseparable from, larger assemblies (e.g., manufactured objects with several components). Current archaeological processes to analyse concealed or nested archaeological material often use destructive techniques. On the other hand, the absence of a real understanding of the internal structure or state of decay of some objects increases the risk that investigation could destroy source material.

CT scan is an imaging technology based on X-rays mostly used for medical purposes. It produces images of the internal structure of the scanned objects with density information about the internal composition. This technology is increasingly used in Cultural Heritage (CH) to obtain images of the internal structure of archaeological material. However, it remains mainly limited to providing 2D images.

We propose a workflow where the CT scan images are used to produce volume and surface 3D data which serve as a basis for new evidence usable by archaeologists. This new evidence can be observed in interactive 3D digital environments or through physical copies of internal elements of the original material, as in [1] and [2].

This workflow has been applied to a block of corroded Iron Age tools discovered in Plumaugat [L3], Brittany, France (Figure 1) during excavations conducted by E. Ah Thon, Inrap. The CT scan of the block revealed an assembly of several blacksmith tools (Figure 2). The resulting

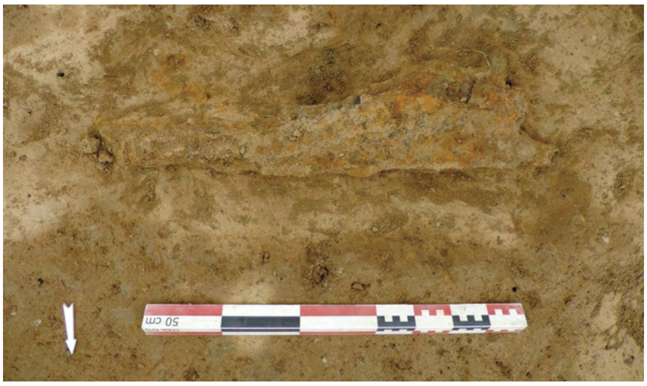


Figure 1: The original block excavated from the site of Plumaugat.

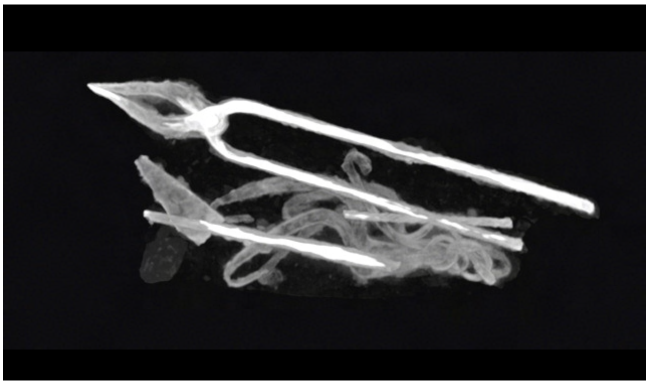


Figure 2: View of the internal structure of the block with CT scan.

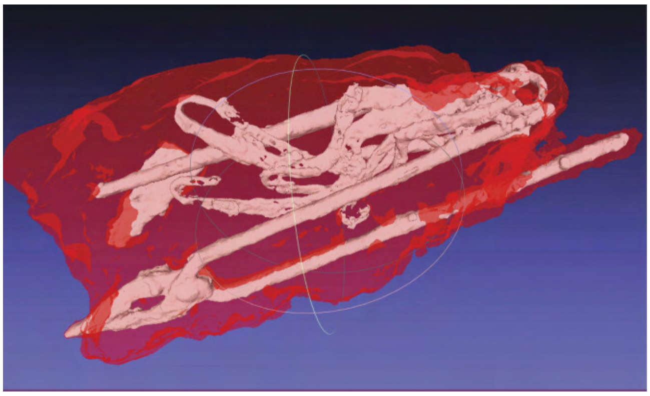


Figure 3: 3D models of the external shape and the internal tools.

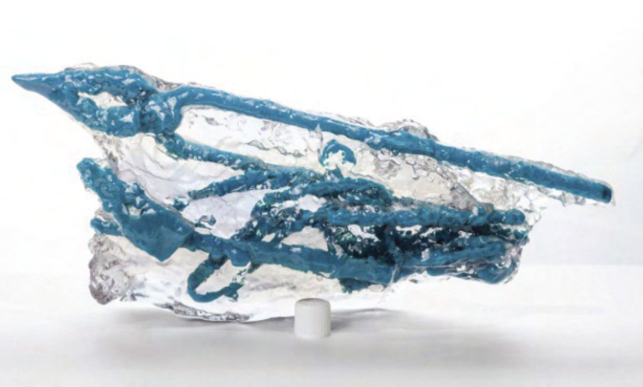


Figure 4: 3D printing of the block.

DICOM data was processed with the Osirix software in order to generate 3D models of the metal tools and of the external shape of the block (Figure 3). These 3D models were then processed and 3D printed with an emerging 3D printing technique mixing coloured and transparent parts (Figure 4).

The resulting object is a 1:1 physical representation of the initial object that gives access to the internal spatial organisation of the components. The tangible medium allows for physical manipulation and simple visualisation to support researchers' analysis, as well as aiding the excavation process as such. Having such representations available offers the possibility of taking immediate precautionary measures before any manual intervention; such representations are also the only tangible medium of context preserved after the excavation of the original material. Furthermore, it is important to note that the workflow process presented here is far quicker than a current excavation and restoration processes as presented in [3].

The workflow presented in this paper is currently developed and extended through collaboration with the Canadian INRS and University of Laval, within the ANR-FRSCQ project INTROSPECT [L4]. This project, which started in January 2017, aims to develop digital interactive introspection methods for archaeological material. These methods will combine CT scan with 3D technologies such as virtual and augmented reality, tangible interactions, and 3D printing.

Thanks to the interdisciplinary collaboration, the INTROSPECT project aims to provide innovative solutions with new usages and tools to allow access to new knowledge for archaeologists and CH practitioners. The project is based on real use cases corresponding to actual archaeological problems. The scientific heart of the project is the systematisation of the relation between the object/artefact, the archaeological context, the digital object, the virtual reconstruction of the archaeological context, as well as the physical copy obtained with 3D printing.

Links:
[L1] www.inrap.fr
[L2] www.inria.fr
[L3] https://kwz.me/hnL
[L4] introspect.info/

References:

[1] T. Nicolas, et al.: "Preservative Approach to Study Encased Archaeological Artefacts, Int. Conf. on Cultural Heritage, LNCS, Vol. 8740, pp. 332–341, 2014.

[2] T. Nicolas, et al.: "Touching and interacting with inaccessible cultural heritage, in Presence: Teleoperators and Virtual Environments", MIT Press, 2015, 24 (3).

[3] T. Nicolas, et al.: "Internal 3D Printing of Intricate Structures", Int. Conf. on Cultural Heritage, LNCS, Vol. 10058 (Part I), pp.432-441, 2016.

Please contact:

Valérie Gouranton
INSA de Rennes/ IRISA-Inria, France,
Valerie.Gouranton@irisa.fr

Théophane Nicolas, Inrap/UMR 8215
Trajectoires, France
theophane.nicolas@inrap.fr

Jean-Baptiste Barreau
Jean-Baptiste Barreau, CNRS/
CReAAH UMR 6566, France
jean-baptiste.barreau@univ-rennes1.fr

Reinterpreting European History Through Technology: The CrossCult Project

by Susana Reboreda Morillo (Universidad de Vigo), Maddalena Bassani (Università degli Studi di Padova) and Ioanna Lykourentzou (Luxembourg Institute of Science and Technology)

The H2020 CrossCult project aims to spur a change in the way Europeans appraise history.

European history is an exciting mesh of interrelated facts and events, crossing countries and cultures. However, historic knowledge is usually presented to the non-specialist public (museum or city visitors) in a siloed, simplistic and localised manner. In CrossCult [L1], an H2020-funded EU project that started in 2016, we aim to change this [1, 2]. With an interdisciplinary consortium of 11 partners, from seven European countries [L2] we are developing technologies to help answer two intrinsically united humanities challenges:

- How can we present historic knowledge to non-specialist audiences in an engaging way?
- How can we further trigger these audiences to reflect, individually and collectively, on European history and its connection to the present?

In an era where unity is more important than ever in Europe, the CrossCult project contributes to the understanding of otherness, and shows the importance of the past to explain the present.

CrossCult is implemented through four pilots, which act as real-world demonstrators of what we aim to achieve. From large museums to small ones, and from indoors to outdoors, each pilot represents a strategically important type of European historical venue.

Pilot 1 – Large multi-thematic venue: Building narratives through personalisation

Pilot 1 takes place in the National Gallery in London. It triggers reflection through personalisation, as it uses the gallery's large collection to offer the visitors personalised stories that highlight the connections among people, places and events across European history, through art. Semantic reasoning, recommender systems and path routing optimisation are employed to ensure that each visitor will be navigated through the conceptually linked exhibits that interest them the most, while avoiding congested spaces as much as possible.

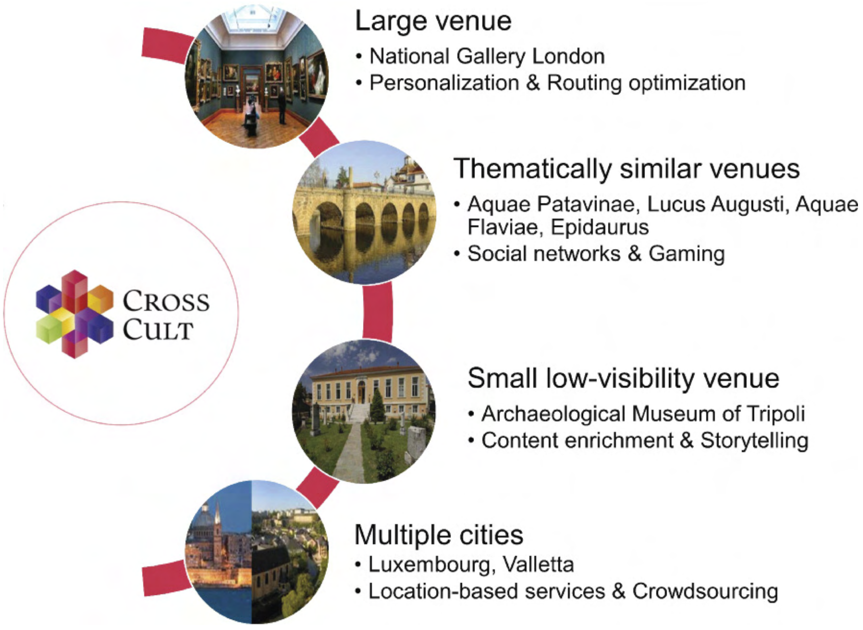

Cross Cult
Large venue
• National Gallery London
• Personalization & Routing optimization
Thematically similar venues
• Aquae Patavinae, Lucus Augusti, Aquae Flaviae, Epidaurus
• Social networks & Gaming
Small low-visibility venue
• Archaeological Museum of Tripoli
• Content enrichment & Storytelling
Multiple cities
• Luxembourg, Valletta
• Location-based services & Crowdsourcing

Figure 1: CrossCult H2020 project – Overview of the four pilots and their supporting technologies.

The improved quality of experience that this combination of technologies offers, balancing in a unique way individual visitor needs with museum-wide objectives, can be extended and customised to serve the needs of various other large venues across Europe.

Pilot 2 – Many venues of similar thematic: Building narratives through social networks and gaming

The second pilot triggers reflection through socialisation. It takes place in four different small venues (Montegrotto Terme/Aquae Patavinae, Italy; Lugo/Lucus Augusti, Spain; Chaves/Aquae Flaviae, Portugal; Epidaurus, Greece), where thermo-mineral water and its use is presented as one of the most important natural resources of the past and the present. Through social networked gaming and storytelling, the pilot connects remote groups of visitors helping them understand the historic connections shared by the venues, in this case on thermalism and water (composition, cults, health treatments, pilgrimages, leisure and daily life).

The focus on multiple venues is strategically important, due to the huge number of small and medium-sized museums around Europe connected through similar themes.

Pilot 3 – Small venue: Building narratives through content enrichment

Pilot 3 takes place in a peripheral, low-profile museum, the Archaeological Museum of Tripoli in Greece. Like many other small museums all over Europe, this venue houses interesting objects, but offers limited informative material and is often overshadowed by larger venues. Content enrichment, in the form of digital narratives, storytelling and social media, backed by psychology techniques, such as empathy increase, are used to create a non-typical visit that goes beyond a traditional object showcase and immerses visitors into life, power structures, and the place of women in antiquity. History, in this respect, becomes relevant to people's lives, enabling them to see the connections between their present and the past, by reflecting on ever-important human issues like religion, mortality and social equality.

The example of pilot 3 can be applied on multiple other small cultural spaces to help raise their profile and take advantage of the latest social, educational, technological developments.

Pilot 4 – Multiple cities: Building narratives through urban informatics and crowdsourcing

European history is intrinsically connected with location; our cities, buildings and streets bear the marks of the people and populations that have inhabited them. Pilot 4 takes place outdoors in two cities, Luxembourg City in Luxembourg and Valletta in Malta, and triggers reflection through urban discovery. Focusing on the topic of migration, past for Malta and present for Luxembourg, and using the technologies of location-based services, urban informatics and crowdsourcing, it invites people to walk the two cities, discover and share stories. Visitors and residents engage in comparative reflection that challenges their perception on topics touched by migration such as identity, quality of life, traditions, integration and sense of belonging.

This pilot has significant potential for the promotion of cultural tourism, as its technologies support the easy integration of new cities to its existing seed city network.

A Living Lab that you are welcome to join

CrossCult implements its vision through a living lab approach: we invite researchers, cultural heritage representatives and stakeholders to co-design with us, take part in our experiments and share their viewpoints in a network of venues and experts across Europe. Get in touch with us at contact@crosscult.eu.

Acknowledgements:
We are grateful to Dr. Angeliki Antoniou, Dr. Evgenia Vasilakaki and the whole CrossCult consortium.

Links:
[L1] http://crosscult.eu/
[L2] https://kwz.me/hnu

References:
[1] I. Lykourentzou et al..”Reflecting on European History with the Help of Technology: The CrossCult Project”, GCH 2016
[2] C. Vassilakis et al..”Interconnecting Objects, Visitors, Sites and (Hi)Stories Across Cultural and Historical Concepts: The CrossCult Project”, EuroMed 2016

Please contact:
Susana Reboreda Morillo
Universidad de Vigo, Spain
+34 988 387 269, rmorillo@uvigo.es

Maddalena Bassani
Università degli Studi di Padova, Italy
+39 329 987 7881,
maddalena.bassani@unipd.it

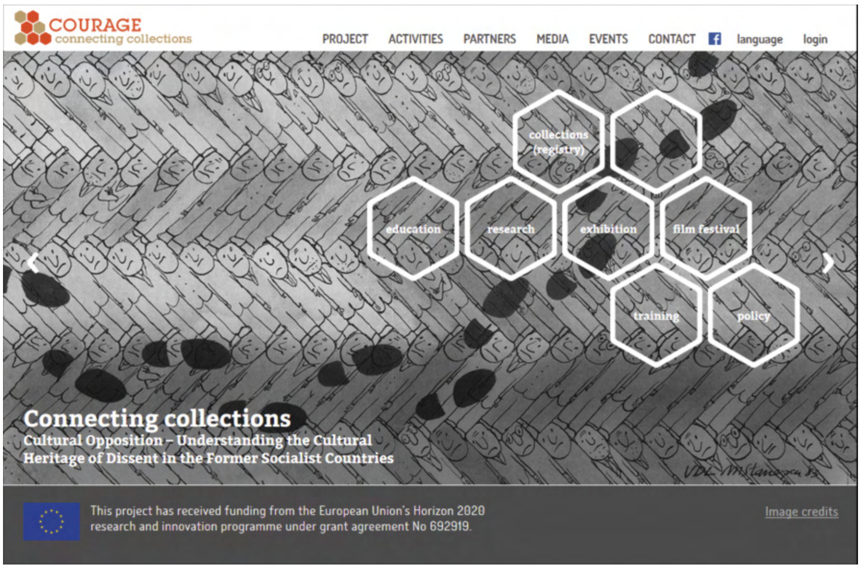
Ioanna Lykourentzou, Luxembourg Institute of Science and Technology
+352 275 888 2703,
ioanna.lykourentzou@list.lu

Cultural Opposition in former European Socialist Countries: Building the COURAGE Registry

by András Micsik, Tamás Felker and Balázs Nász (MTA SZTAKI)

The COURAGE project is exploring the methods for cultural opposition in the socialist era (cc. 1950-1990). We are building a database of historic collections, persons, groups, events and sample collection items using a fully linked data solution with data stored in an RDF triple store. The registry will be used to create virtual and real exhibitions and learning material, and will also serve as a basis for further narratives and digital humanities (DH) research.

COURAGE addresses the role of the collections in defining what “cultural opposition” means in order to facilitate a better understanding of how dissent and criticism were possible in the former socialist regimes of Eastern Europe.

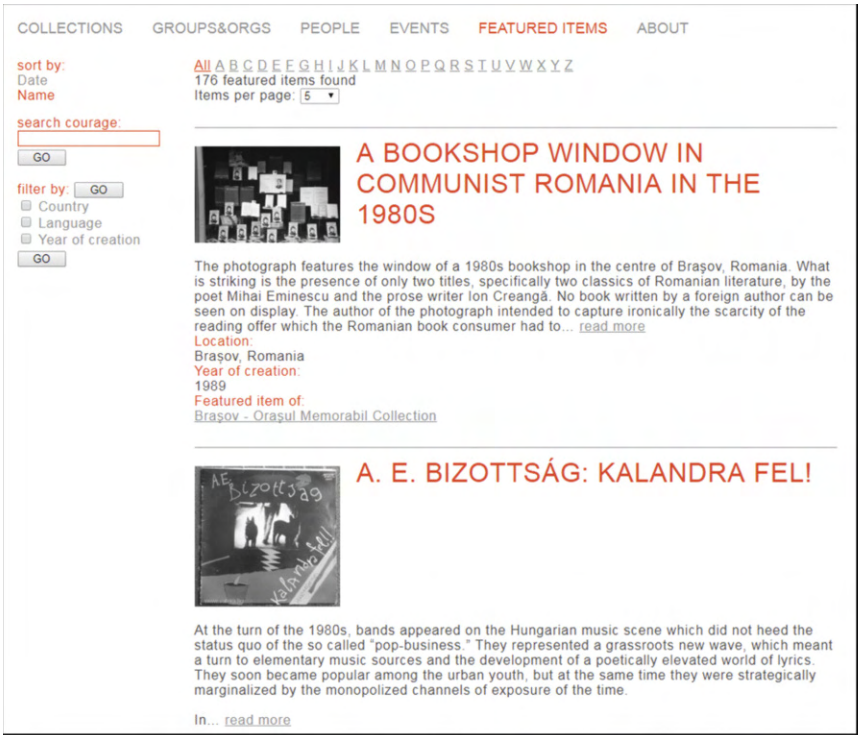


Connecting collections

Cultural Opposition – Understanding the Cultural Heritage of Dissent in the Former Socialist Countries

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 692919.

Figure 1: The COURAGE project portal.



176 featured items found

Items per page: 5

A BOOKSHOP WINDOW IN COMMUNIST ROMANIA IN THE 1980S

The photograph features the window of a 1980s bookshop in the centre of Brașov, Romania. What is striking is the presence of only two titles, specifically two classics of Romanian literature, by the poet Mihai Eminescu and the prose writer Ion Creangă. No book written by a foreign author can be seen on display. The author of the photograph intended to capture ironically the scarcity of the reading offer which the Romanian book consumer had to... read more

Location:
Brașov, Romania

Year of creation:
1989

Featured item of:
Brașov - Orașul Memorabil Collection

A. E. BIZOTTSÁG: KALANDRA FEL!

At the turn of the 1980s, bands appeared on the Hungarian music scene which did not heed the status quo of the so called “pop-business.” They represented a grassroots new wave, which meant a turn to elementary music sources and the development of a poetically elevated world of lyrics. They soon became popular among the urban youth, but at the same time they were strategically marginalized by the monopolized channels of exposure of the time.

In... read more

Figure 2: List of sample items from the registry.

The type of opposition we want to discover was largely evident in the culture and lifestyle under the socialist era, and include activities such as alternative music, alternative fine arts, folk dance clubs and religious movements. The COURAGE project [1] [L1] (within the EU H2020 framework) is creating a comprehensive online database (digital registry) of existing but scattered collections on the histories and forms of cultural opposition in the former socialist countries, thereby making them more accessible. It will analyse these collections in their broader social, political and cultural contexts. The general aim of this analysis is to allow for the expanded outreach and increased impact of the collections by assessing the historical origins and legacies of various forms of cultural opposition. Our research team aims to explore the genesis and trajectories of private and public collections on cultural opposition movements, the political and social roles and uses of the collections before 1989 and since, the role of exiles in supporting these collections, etc.

The project team contains institutes or faculties of history or sociology from Bulgaria, Croatia, Czech Republic, Germany, Hungary, Ireland, Lithuania, Poland, Romania, Slovakia and the United Kingdom. The role of IT support is assigned to MTA SZTAKI, Hungary.

Research results will be made available in multiple forms. The online education material will bring to light the hidden and lesser-known cultural life of the former socialist countries and will facilitate teaching and learning about the period. The exhibition will give access to the hitherto less known masterpieces of cultural opposition as well as the interesting lifestyles of its members through fascinating visual footage and freshly disclosed archival documents. Policy documents will help decision makers and funders to recognise the social impact and use of the collections of cultural opposition, and will provide a forum to discuss projects of common interest. COURAGE will also organise a documentary film festival. We want to encourage younger generations to care about the past, and to appreciate the importance of cultural opposition.

Based on the requirements set by historians and sociologists, MTA SZTAKI installed and customised a linked data platform, where data for all former socialist countries can be entered by researchers involved in the project. The linked data approach was a natural choice given the importance of capturing the connections between all the researched players, events and collections. The linked data approach is not yet widely known and appreciated in the fields of arts and social science, but our researchers quickly realised its advantages and learned about its different way of data representation. The editor team benefits from the easy traversal and grouping of linked entities and the multilingual text handling, while the administrators can very easily extend or change the data model on-the-fly. Editors use the Vitro platform with Jena triple store for editing, which is connected to a set of WordPress portals for public presentation.

Among the many IT challenges, first we had to solve the representation of timed properties, for example when a collection was owned by a person and then donated to a museum. Secondly, we had to implement a kind of simplified authorisation for triple editing (the option “Everything is open” was not acceptable for the community). Thirdly, a quality assurance workflow was also developed in-house. Entities in the registry go through a seven-step workflow containing local and central quality checks and English proofreading. The result is what we can call a knowledge graph connecting collections, events and participants of cultural opposition. Part of this graph is already published and browseable on the project portal under the registry link. The schema of the registry is called the COURAGE Ontology, and it is made available as an OWL file. Owing to our special requirements, we could hardly rely on existing ontologies, but in the future, we aim to map our ontology to various widely used metadata formats.

As our current status in rough numbers, we have 300 collections with 500 featured items, 700 persons, 300 groups or organisations and 400 events in the registry. With 80 researchers editing the registry we are not yet halfway through our planned work. We hope that by the end of the project we can build an almost complete encyclopaedia of cultural opposition in former socialist countries in the form of a knowledge graph.

Link:
Project home page: http://cultural-opposition.eu/

Reference:

[1] COURAGE: Understanding the Cultural Heritage of Dissent in Former Socialist Countries. Understanding Europe – Promoting the European Public and Cultural Space, 17 October 2016, Brussels.

Please contact:

András Micsik
MTA SZTAKI, Hungary
+36 1 279 6248
andras.micsik@sztaki.mta.hu

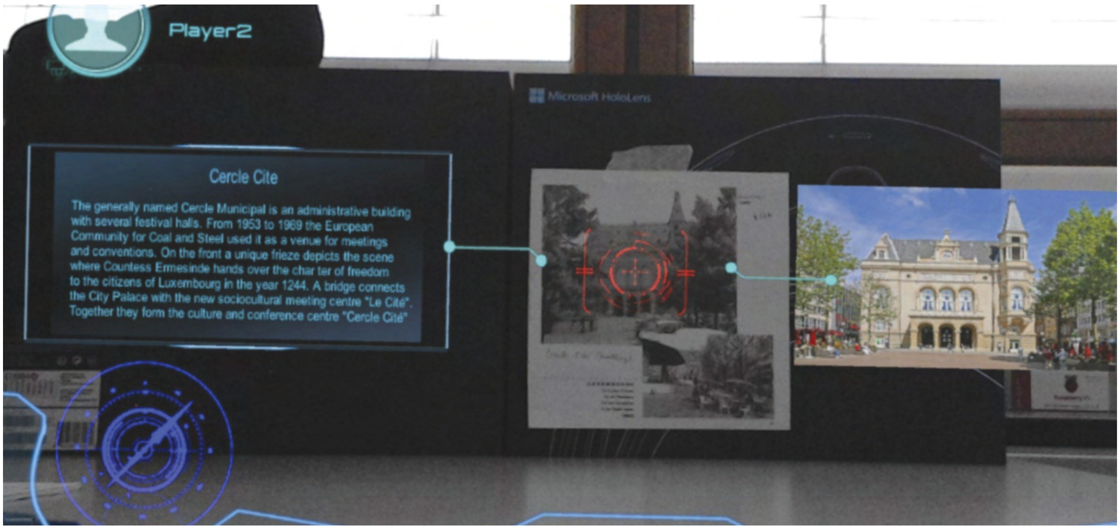
Locale, an Environment-Aware Storytelling Framework Relying on Augmented Reality

by Thomas Tamisier, Irene Gironacci and Roderick McCall (Luxembourg Institute of Science and Technology)

The Locale project proposes a vision of location-aware digital storytelling empowered by a combination of technologies including data mining, information visualisation and augmented reality. The approach is tested through pilot contributors who share their experiences, stories and testimonies of Luxembourg since the end of World War II.

Worldwide, and especially in the European Union, information technology is playing an increasingly important role in enabling the efficient use and preservation of cultural heritage. The Locale project looks at the preservation of immaterial cultural heritage from the perspective of tools and techniques to support the connection of different stories. In fact, navigating complex information spaces remains a challenge despite continuous improvements in search and retrieval processes, especially when the information refers to overviews or miscellaneous references, such as when connecting particular stories and places. The Locale project is based on a collaborative mobile and web-based platform with a focus on location-based storytelling for sharing testimonies and multi-media historical heritage content relating to the period of 1945-1960 in Luxembourg and the Greater Region: from the end of WWII to the dawn of Europe, in the context of their respective 70th (2015) and 60th (2017: European Economic Community) anniversaries.

In addition, Locale intends to foster the sharing of personal historical accounts that might not be included in standard historical literature. The platform offers dedicated functionalities for exploring multidimensional data using various human analysis and data mining strategies, based on metadata, tags, attributes entered by the user, and browsing history (e.g., connections between a place and queries about a given historical fact). In particular, Locale draws from collaborative visualisation, allowing users to share views of the same content with different focuses, and providing an intuitive way of sharing content [1]. The main operational features include, on the one hand, text and data mining functionalities for updating knowledge and trig-



Cercle Cite

The generally named Cercle Municipal is an administrative building with several festival halls. From 1953 to 1969 the European Community for Coal and Steel used it as a venue for meetings and conventions. On the front a unique frieze depicts the scene where Countess Ermesinde hands over the char ter of freedom to the citizens of Luxembourg in the year 1244. A bridge connects the City Palace with the new sociocultural meeting centre "Le Cité". Together they form the culture and conference centre "Cercle Cité"

Figure 1: Augmented reality in Locale: Detection and tracking of a target image of a building.

gering selective actions supporting the interaction between platform users and, on the other hand, different modalities of exploring and editing stories with a view to enhancing the spatial dimension and feeling of flow and immersion by using a desktop, a mobile device, or interactive augmented reality equipment.

Location-based storytelling requires that people feel immersed in the experience, and perhaps even feel part of it (i.e. present). For this to be achieved it is essential that any system makes the user feel a sense of space and more importantly place. Space relates to the physical properties of the environment, for example street layouts, buildings and perhaps smaller aspects such as benches. In contrast, a sense of place is when a space becomes infused by different meanings. A sense of place arises out of the blending of aspects such as sense of self, other people and activities in the space [2]. Earlier storytelling applications using mobile phones explored this concept [3] and in Locale this is further enhanced through the co-construction between different physical and narrative elements, combined with augmented reality technology.

By filtering physical environments through ad-hoc additional features and augmented reality, Locale has been tested with different use-cases and proven to bring powerful support in revisiting a scene of a story across time and users' perspectives. For example, a route can be created which contains multiple stopping points, as the user walks along they can listen to stories about places, people and events at different locations. Furthermore, if there are many layers or stories at a specific location about a particular person or event this may give the user a stronger sense of history and importance and ultimately shape their understanding of that place. In this regard, Locale provides specifically new ways of interaction through the use of an AR headset, a new type of non-command user interface able to track user movements and use them to create UI elements the user can interact with. Additional visualisations provide: indications of the degree of agreement/disagreement between sources of information available and links between related information. Figure 1 shows the test of Locale in the Virtual Reality laboratory of LIST: a real grayscale image (centre) is seen through a Microsoft HoloLens headset, which triggers augmented content consisting of a current image of the building (right) and related information (left). The overall picture is an interactive and immersive storytelling experience where the user can interact with the contents of the story (notably images and 3D models) in a simple and natural way, for example using gestures or voice.

As a whole, Locale provides an operational framework for location aware and collaborative storytelling that focuses on three main challenges identified in the literature. First, it encourages people to structure stories in ways that support their perception of place and sense of presence. Second, it enables the linking of content and mining of related data to improve how people can navigate within stories and spaces as well as provide people with easier ways to see and interact with the rich content. Finally, it explores novel interface techniques that are designed to present complex information but avoid information overload. As a primary impact, the operational framework achieved through the project will help to explore how technologies coupled with an environment-aware setting can help to bridge the digital divide between users of different ages and backgrounds.

Acknowledgments:
The Locale project is funded by Luxembourg's Fonds National de la Recherche under the Core grant C12/IS/3986199.

Links:
http://list.lu
https://kwz.me/hNN

References:
[1] P.Carvalho: "Using Visualization Techniques to acquire a better understanding of Storytelling", Proc. of DATA conference, 2017.
[2] P. Gustafson: "Meanings of place: everyday experience and theoretical conceptualizations", Environmental Psychology. Elsevier, 21.5-16, 2001.
[3] R. McCall: "Mobile phones, Sub-Culture and Presence", in Proc. of the workshop on Mobile Spatial Interaction at ACM Conference on Human Factors in Computing Systems (CHI), 2011.

Please contact:
Thomas Tamisier, Luxembourg Institute of Science and Technology
thomas.tamisier@list.lu

The Biennale 4D Project

by Kathrin Koebel, Doris Agotai, Stefan Arisona (FHNW) and Matthias Oberli (SIK-ISEA)

Virtual reality (VR) reconstruction offers a new interactive way to explore the archives of the Swiss Pavilion at the "Biennale di Venezia" Art Exhibition.

The Swiss pavilion at "Biennale di Venezia" offers a platform for national artists to exhibit their work. This well-known white cube showcases the changes in contemporary Swiss art from the early 50s to the present day. The aim of the "Biennale 4D" is to make the archives of the past bi-annual art exhibitions more comprehensible by creating an interactive explorative environment using innovative virtual reality (VR) technology. «Biennale 4D» poses multiple challenges including visualisation of historic content and its documentation, dealing with the heterogeneity and incompleteness of archives, interaction design and interaction mapping in VR space, integration of metadata as well as realising a virtual reality experience for the public space with current VR technology.

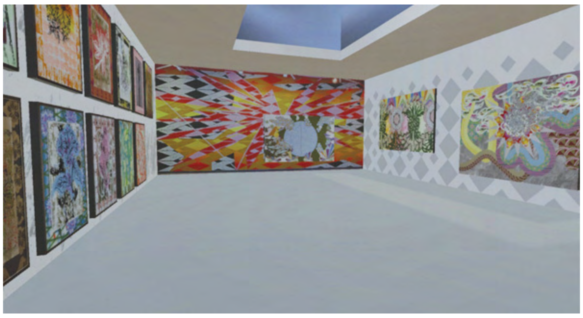


Figure 1: Screenshot of the Biennale 4D prototype, showing the virtual reconstruction of the 2007 art exhibition in the "Malerei" hall of the Swiss pavilion.

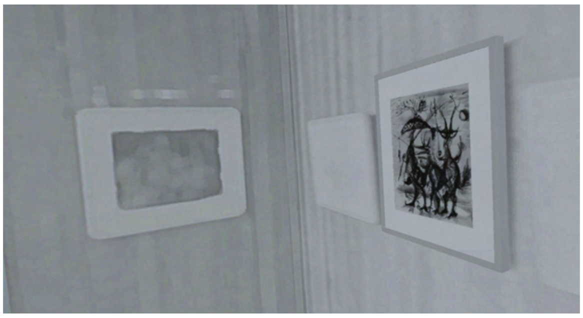


Figure 2: Screenshot of the Biennale 4D prototype, displaying the chosen aesthetics for the visualisation of the historical content. The original design of the pavilion was reduced and blur has been added to the wall textures in order to intuitively guide the user's focus to the documented art works.

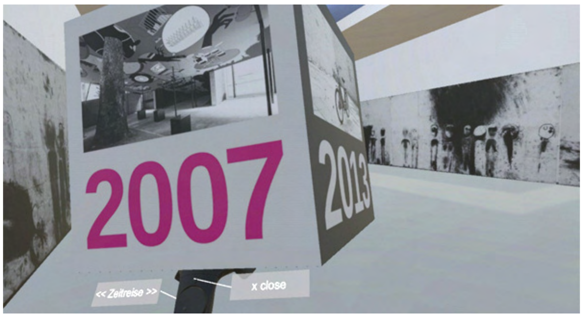


Figure 3: Screenshot showing the time machine object which was created to allow the user to interact with the time dimension.

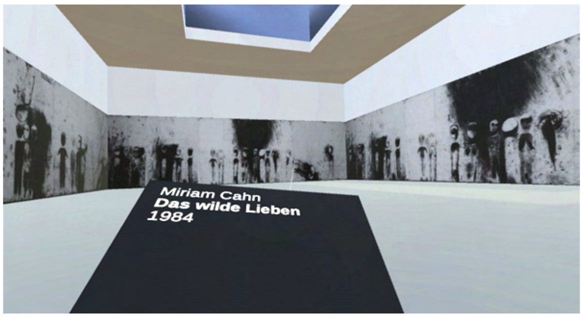


Figure 4: Screenshot of the information guide that allows interaction with the metadata of the art work.

Design and development of the reconstructed exhibition environment

A pilot application was created by the Institute of 4D Technologies of the University of Applied Sciences and Arts Northwestern Switzerland FHNW and Swiss Institute for Art Research SIK-ISEA [1] to test the concept of the proposed VR experience. This functional prototype allows users to explore the pavilion and displayed art works of different epochs with the HTC Vive VR headset and hand controllers. It contains a 3D model of the pavilion based on the original design by Bruno Giacometti and a concept for visualisation of the exhibition content and its documentation. Exhibition samples (see Figure 1) of various documentation levels are included – for example, "thoroughly documented", "fragmentarily documented" – as well as experimental art works, such as video installations. The current release of the prototype showcases exemplary portions of the selected exhibitions in the years 1952, 1984, 2007 and 2013. The development of a consistent visual language for the heterogeneous work posed a challenge. Numerous experiments were made to discover a suitable visualisation style (see Figure 2) that guides the user's focus away from the building to the artworks and their documentation. These experiments included variables such as the degree of abstraction, deliberate lack of definition and level of chromaticity, and lessons from the field of archaeology have been applied for the handling of incomplete data.

Interaction with time, space and metadata

The prototype allows interaction with three dimensions which had to be reduced onto the two hand controllers available to the user. In this initial prototype one hand is assigned to time travel and the other hand is designated for spatial movement and interaction with the objects. The user is able to travel intuitively through time by means of interaction with the time machine object (see Figure 3). This three-dimensional item offers affordances to the user about the exhibition content of the years he's passing by as he or she travels through time, navigating to the scene of the desired year. For spatial movement the application contains a navigation concept that allows the user to move within the virtual room either by position tracking of the user's movement in the physical space or via teleportation. It provides basic haptic feedback in case of collisions. Other forms of spatial navigation were considered (e.g., a guided tour), however testing revealed a correlation between the user's degree of freedom regarding movement and the perceived user experience. In addition, an information guide in the form of a virtual booklet offers metadata corresponding to the objects on display (see Figure 4). This supplementary information can be accessed by pointing with a laser ray towards the desired item. And, interactive hotspots have been added to the application to show additional material, such as archive photos.

Conclusion and future work

Biennale 4D poses a unique challenge as it combines a virtual art exhibition experience with archive functionalities through the use of virtual reality technology. This synthesis allows new approaches to exhibition reconstruction and the conclusive incorporation of historical material in this experimental virtual space. The curative work intertwines three layers of materials: histori-cal content (original artwork), its doc-umentation (archive photos and other artefacts) and the virtual room (mapping space). In particular the handling of the documentation layer leaves much room for interpretation and exploration. Furthermore, the nature of this application field requires thoughtful examination of aspects like substantiality, aesthetics as well as the concept of time.

Some other areas of focus for further work include further elaboration on the aesthetics of the visualisation including spatial design of the surrounds, improving the storytelling and developing a concept to present the application and allow its usage in the public space. Ongoing work will focus on offering access to the full content of the Biennale archives in an even more interactive and immersive way and letting a wider audience experience this valuable portion of Swiss art and cultural history.

Links:
www.biennale-venezia.ch
www.fhnw.ch/technik/i4ds

Reference:

[1] K. Koebel, O. Kaufmann, "Biennale 4D – Erschaffung einer Virtual Reality Experience zur Exploration der Archivbestände des Schweizer Pavillons an der Biennale Venezia", https://kwz.me/hLd.

Please contact:

Kathrin Koebel
Fachhochschule Nordwestschweiz, Switzerland
kathrin.koebel@students.fhnw.ch

The Clavius Correspondence: From Digitization to Visual Exploration of Knowledge

by Matteo Abrate, Angelica Lo Duca, Andrea Marchetti (IIT-CNR)

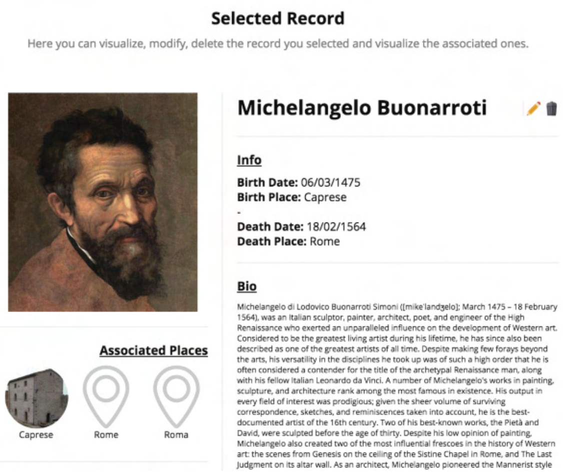
The "Clavius on the Web Project" [L1] is an initiative involving the National Research Council in Pisa and the Historical Archives of the Pontifical University in Rome. The project aims to create a web platform for the input, analysis and visualisation of digital objects, i.e., letters sent to Christopher Clavius, a famous scientist of the 17th Century.

In the field of digital humanities, cultural assets can be valued and preserved at different levels, and whether or not an object is considered a knowledge resource depends on its peculiarity and richness. Within the Clavius on the Web Project we mainly consider two kinds of knowledge resources: contextual resources associated to digitised documents and manual annotations of cultural assets. We implemented a different software for each of these resource types: the Web Metadata Editor for contextual resources and the Knowledge Atlas to support manual annotation.

Semi-automated enrichment of catalogues: the Web Metadata Editor

In recent years, a great effort has been made within the field of digital humanities to digitise documents and catalogues in different formats, such as PDF, XML, plain texts and images. These documents are often stored either in digital libraries or big digital repositories in the form of books and catalogues. The process of cataloguing also requires the creation of a knowledge base, which contains contextual resources associated to documents of the catalogue, such as the authors of the documents and places where documents were written. Information contained in the knowledge base can be used to enrich document details, such as metadata associated to documents. Most of the existing tools for catalogue creation allow the knowledge base to be built manually. This process is often tedious, because it requires known information about a document, such as the author's name and date of birth, to be edited. It is also a repetitive process because many documents are written by the same author and in the same place thus requiring the same information to be written twice or more. In general there are three main disadvantages of this manual effort, compared to an automated system: (i) there is a higher probability of introducing errors; (ii) the process is slower; (iii) the entered information is isolated, i.e., not connected to the rest of the web.

In order to mitigate these disadvantages in the context of the project, we developed the Web Metadata Editor (WeME) [1], a user-friendly web application for building a knowledge base associated to a catalogue of digital documents. Figure 1 shows a snapshot of WeME. While the application is envisaged for archivists/librarians, it may also be useful for others. WeME helps archivists to



Selected Record

Here you can visualize, modify, delete the record you selected and visualize the associated ones.

Michelangelo Buonarroti

Info

Birth Date: 06/03/1475
Birth Place: Caprese
-
Death Date: 18/02/1564
Death Place: Rome

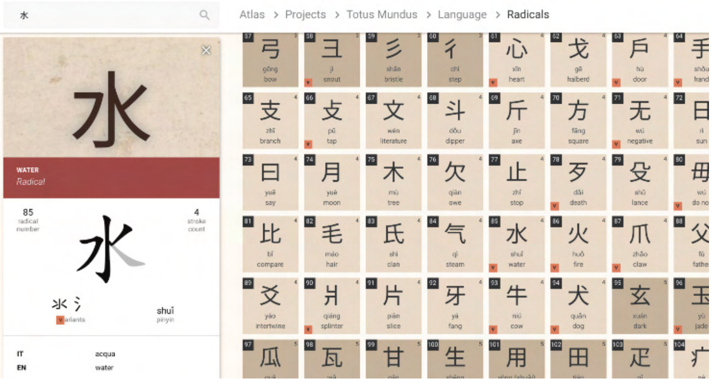
Bio

Michelangelo di Lodovico Buonarroti Simoni ([mikeˈlandʒelo]; March 1475 – 18 February 1564), was an Italian sculptor, painter, architect, poet, and engineer of the High Renaissance who exerted an unparalleled influence on the development of Western art. Considered to be the greatest living artist during his lifetime, he has since also been described as one of the greatest artists of all time. Despite making few forays beyond the arts, his versatility in the disciplines he took up was of such a high order that he is often considered a contender for the title of the archetypal Renaissance man, along with his fellow Italian Leonardo da Vinci. A number of Michelangelo's works in painting, sculpture, and architecture rank among the most famous in existence. His output in every field of interest was prodigious; given the sheer volume of surviving correspondence, sketches, and reminiscences taken into account, he is the best-documented artist of the 16th century. Two of his best-known works, the Pietà and David, were sculpted before the age of thirty. Despite his low opinion of painting, Michelangelo also created two of the most influential frescoes in the history of Western art: the scenes from Genesis on the ceiling of the Sistine Chapel in Rome, and The Last Judgment on its altar wall. As an architect, Michelangelo pioneered the Mannerist style

Associated Places

Caprese Rome Roma

Figure 1: A snapshot of WeME.



水

Atlas > Projects > Totus Mundus > Language > Radicals

水

WATER
Radical

85 radical number

4 stroke count

水

氺 氵 V ariants

shuǐ pinyin

IT acqua
EN water

57 弓 3 gōng bow	58 彐 3 V jì snout	59 彡 3 shān bristle	60 彳 3 chì step	61 心 4 V xīn heart	62 戈 4 gē halberd	63 戶 4 V hù door	64 手 V shǒu hand
65 支 4 zhī branch	66 攴 4 V pū tap	67 文 4 wén literature	68 斗 4 dǒu dipper	69 斤 4 jīn axe	70 方 4 fāng square	71 无 4 V wú negative	72 日 rì sun
73 曰 4 yuē say	74 月 4 yuè moon	75 木 4 mù tree	76 欠 4 qiàn owe	77 止 4 zhǐ stop	78 歹 4 V dǎi death	79 殳 4 shū lance	80 毋 V wú do no
81 比 4 bǐ compare	82 毛 4 máo hair	83 氏 4 shì clan	84 气 4 qì steam	85 水 4 V shuǐ water	86 火 4 V huǒ fire	87 爪 4 V zhǎo claw	88 父 fù fathe
89 爻 4 yáo intertwine	90 爿 4 V qiáng splinter	91 片 4 piàn slice	92 牙 4 yá fang	93 牛 4 V niú cow	94 犬 4 V quǎn dog	95 玄 5 xuán dark	96 玉 V yù jade
97 瓜 5	98 瓦 5	99 甘 5	100 生 5	101 用 5	102 田 5	103 疋 5	104 疒

Figure 2: A screenshot of Knowledge Atlas showing information about a Chinese radical.

enrich their catalogues with resources extracted from two kinds of web sources: Linked Data (DBpedia GeoNames) and traditional web sources (VIAF).

WeME, which is published as an open source software on GitHub [L2], was tested by 31 users, 45.2% of whom were female and 54.8% male. Almost half (48.8%) of the testers were older than 35 years. In a scale ranging from poor (1) to excellent (5), the scores given to WeME were: 4 (by 50% of testers), 5 (by 20%), 3 (by 23.3%) and 1 or 2 by the remaining testers. Tests indicate that WeME is a promising solution for reducing the repetitive, often tedious work of archivists.

Visual exploration of knowledge: Knowledge Atlas

Creating a catalogue and digitising assets from the archive, while fundamental, soon proved insufficient to convey the richness of the knowledge within the archive. We thus started the development of Knowledge Atlas [L3], a user interface designed with the following principles:

1) Visualisation and interaction – Content presentation should take advantage of the visual expressiveness and interactivity of modern web technologies. For example, the interactive recreation of volvellae, instruments made of rotating wheels of paper, which enables scholars to investigate their purpose without risking damage to the original artefacts.
2) Depth and detail – Content presentation should explicitly show many different layers of analysis and highlight interesting points. For example, a representation of the Kunyu Wanguo Quantu Chinese map is displayed with highlighted toponyms and cartouches, with which the user can interact to read Chinese transcriptions and Italian translations. Each text can then be further explored to access information about specific Chinese graphemes (Figure 2).
3) Context and links – The main content should be complemented with contextual information. Such presentations should enable navigation from content to context and vice versa. For example, senders, recipients and cited historical characters related to the figure of Christophorus Clavius [2] constitute a graph of correspondents, which can be navigated by itself or starting from a specific passage of a letter. Other examples include: the set of evolving mathematical or astronomical conceptualisations and the related lexical terms in Latin, Italian or Greek [3].
4) Divulgation and correctness – The presentation of content should feature specific design choices aimed at capturing the attention of students or casual readers, by leveraging powerful visual languages and a systematic organisation. The power of providing a structured and intuitive overview should be used as a gateway to lead to the most complete and correct information available. As an example, the representation of the context of the aforementioned letters by Galilei about the model of our solar system is a visual diagram making use of a distorted scale. The actual distances and sizes of the objects can be appreciated by accessing detailed data.

Our current platform, in active and open development on Github, implements this design without being too tied to the specificity of the content. By following this methodology, we believe that our software will prove to be useful in contexts beyond the projects discussed here. We are already experimenting with other content, such as more maps, books, and paintings, but also with models of buildings and complex structured data about the internet.

Links:
[L1] http://claviusontheweb.it
[L2] https://kwz.me/hkj
[L3] http://atlasofknowledge.it

References:

[1] A. Lo Duca et. al.: "Web Metadata Editor: a Web Application to Build a Knowledge Base Based on Linked Data", Third Int. Workshop on Semantic Web for Scientific Heritage, Portoroz, Slovenia, 2017.
[2] M. Abrate et al.: "The Clavius on the Web Project: Digitization, Annotation and Visualization of Early Modern Manuscripts", AIUCD Annual Conference, 2014.
[3] S.Piccini et al.: "When Traditional Ontologies are not Enough: Modelling and Visualizing Dynamic Ontologies in Semantic-Based Access to Texts", DH 2016.

Please contact

Andrea Marchetti, IIT-CNR, Italy
andrea.marchetti@iit.cnr.it

Service-oriented Mobile Social Networking

by Stefano Chessa and Michele Girolami (ISTI-CNR)

The Wireless Network Laboratory at ISTI-CNR studies how to exploit mobility and sociality of people in a mobile social network in order to advertise and discover the services provided by devices. This problem is commonly referred to as service discovery. An efficient strategy for advertising to other devices about the existence of new or existing services is proposed alongside a strategy for discovering a specific service.

Mobile social networking is made up of two key-ingredients: people and the smart devices carried by them [1]. First, people follow a social rather than a random mobility pattern. People tend to spend time together with people who have similar interests [2]. Second, devices are designed with advanced hardware and software capabilities, and they can communicate among themselves based on short-range wireless interfaces, thus forming a mobile network. Indeed, they are equipped with different types of resources such as sensing units and several computational capabilities. Such resources can be seen as exploitable services that can be shared with other devices. When the service-oriented approach is applied in a mobile social context, common application problems have to be rethought for this new challenging perspective.

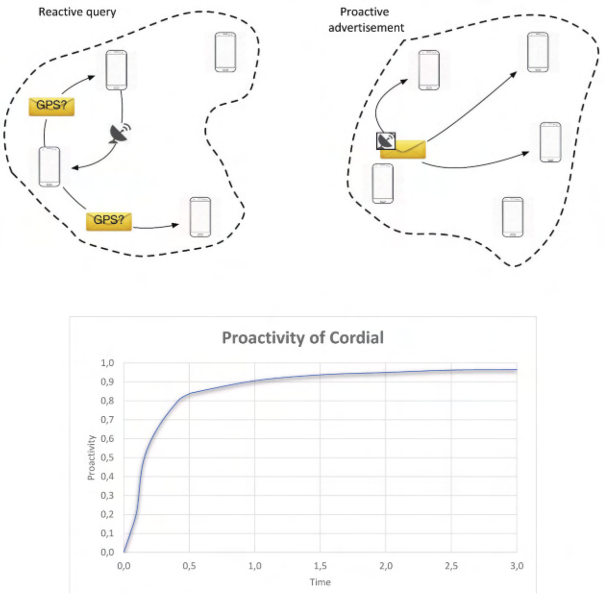


Figure 1: The Reactive and Proactive strategies and Accuracy and Proactivity metrics.

The goal of our research is to study how to exploit mobility and sociality of people in a mobile social network in order to advertise and discover the services provided by devices. This problem is commonly referred to as service discovery.

We investigate the service discovery problem, and in particular the possibility of advertising and discovering services offered by devices in the mobile social network. We propose an efficient strategy for advertising new or existing services to other devices and, at the same time, a strategy for discovering a specific service. The algorithm we have designed is referred to as CORDIAL (Collaborative Service Discovery Algorithm) alluding to the opportunities arising from people close to us.

CORDIAL adopts both a reactive and a proactive mechanism. The reactive mechanism is used to diffuse a service query and implements the query forwarding strategy. This strategy is executed by a device as soon as the user carrying it needs to access a specific service. Consider the following scenario where a user needs to tag a picture with GPS coordinates. The user's device does not provide this functionality, but she can discover devices in proximity that do offer such a service. Once the user finds the required service, in this case the GPS, she can invoke it (see Figure 1).

The basic idea of the reactive phase is to select those devices in range with interests similar to the query sent by the user. Moreover, our strategy gives priority to those devices that are able to interact with other devices (e.g., strangers) that can potentially answer to the service query.

The proactive mechanism of CORDIAL is required to proactively diffuse queries not yet answered (namely the pending queries) and to advertise the existence of services. In this case, the goal is to maximise the diffusion of the messages (queries and service advertisements) and, at the same time, reduce message duplication.

The outcome of this research was the ability to reproduce mobile social networks by using real-world mobility traces. Such traces do a good job of reproducing human mobility in open outdoor environments such as a university campus or a city district. Among all the metrics analysed, we gave particular emphasis to the capacity of the discovery algorithm to store service advertisements of interest for the device, by avoiding the submission of queries (the proactivity metric). Moreover, we analysed the capacity of the algorithm to store only those service advertisements of interest for a device. Such a policy avoids storing off-topic advertisements (we measured such behaviour with the accuracy metric). CORDIAL also obtains optimal values of both metrics (proactivity and accuracy) when compared with other social-aware discovery strategies.

We thus investigated the possibility of exploiting, in an opportunistic fashion, the increasing number of resources offered by smart or low-power devices in mobile social networks. Future work includes the possibility of using CORDIAL in both participatory and opportunistic crowd sensing scenarios as well as using CORDIAL combined together with off-loading computing techniques.

References:

[1] M. Girolami, S. Chessa, A. Caruso: "On Service Discovery in Mobile Social Networks: Survey and Perspectives", Computer Networks, 88, (2015):51-71, DOI 10.1016/j.comnet.2015.06.006

[2] M. McPherson, L. S. Lovin, J. M. Cook, "Birds of a Feather: Homophily in Social Networks", Annual Review of Sociology, 27: 415–444, 2001.

Please contact:

Michele Girolami, ISTI-CNR, Italy
Michele.girolami@isti.cnr.it

Collaboration Spotting: A Visual Analytics Platform to Assist Knowledge Discovery

by Adam Agocs, Dimitris Dardanis, Richard Forster, Jean-Marie Le Goff, Xavier Ouvrard and André Rattinger (CERN)

Collaboration Spotting (CS) is a visualisation and navigation platform for large and complex datasets. It uses graphs and semantic and structural data abstraction techniques to assist domain experts in creating knowledge out of big data.

Valuable knowledge, which can help to solve a range of complex problems, may be created from the extremely large amount of data generated by computers and the internet of things. To achieve this, we rely on sophisticated cognitive tools whose efficacy strongly depends on the delicate interplay between domain experts and data scientists. Domain experts trust data scientists to deliver the tools they need, while the effectiveness of these tools in supporting knowledge creation essentially depends on the information in the hands of domain experts. In other words, creating knowledge essentially relies on the capability of combining domain specific semantic information with concepts extracted out of the data and visualising the resulting networks.

Storing large networks in a flexible and scalable manner calls for graphs: mathematical objects that hold information in nodes representing data instances of particular categories – called facets – and in relationships characterising the network interconnectivity. Facets and relationships embody the network schema, a semantic abstraction of the network content and structure.

Enhancing the cognitive insight of humans into the understanding of the data calls for visual analytics: the science of analytical reasoning supported by interactive visual interfaces that combines the power of visual perception with high performance computing.

Collaboration Spotting (CS) is a data-driven platform based on open source software packages that uses visual analytics concepts and advanced graph processing techniques to provide a flexible environment for domain experts to run their analysis. CS is domain independent and fully customisable. It gives data scientists the capability of building multi-faceted networks out of multiple and heterogeneous data sources and domain experts the ability to specify different perspectives for conducting their analysis by means of network schemas. Individual analysis outputs are visualised in graphs using node-link representations where node size, colour and shape, and relationships highlight the network contents and structures. Output graphs are perspective specific. They represent faceted views of the network under study articulated around part or all of its content. A sophisticated navigation system enables users to graphically interact with individual output graphs and reach other facets of the network. Advanced structural abstraction techniques tailored to graph complexity and size give domain experts a visual

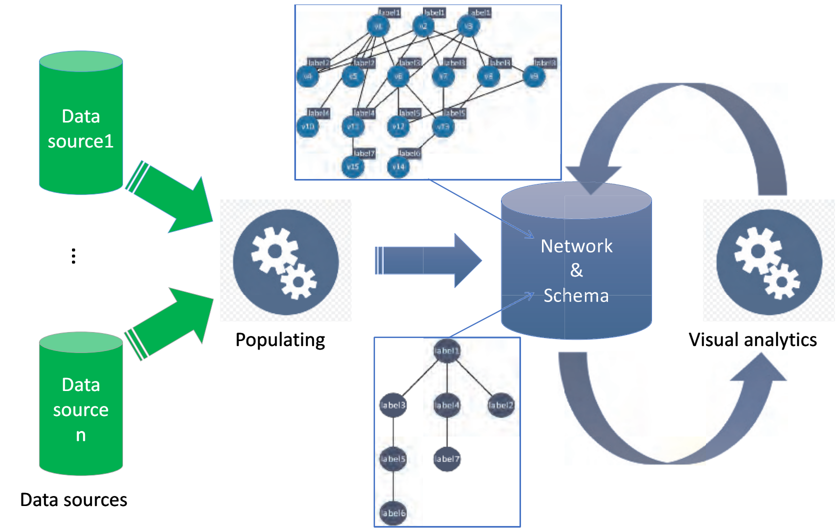


Figure 1: The Collaboration Spotting conceptual framework. After populating the database, domain experts analyse the network content with the help of the schema.

access to fine structures and particularities, such as communities and outliers in a clear manner even for reasonably large graphs. Applying these techniques in hierarchies of graphs provides visual access to larger output graphs at the cost of a loss of semantic and structural information.

Responsiveness when servicing users is an essential aspect of visual analytics. To this end, the CS team has paid particular attention to optimising all the time-critical aspects of the platform. Networks are stored in graph databases that support efficient and flexible query mechanisms. The computing of output graph structures with optimized visual perception is done using scalable clustering and rendering algorithms that run on large computing infrastructures and open source cloud computing software.

CERN has developed Collaboration Spotting with the initial aim of providing the particle physics community with intelligence on academia and industry players active around key technologies with a view to fostering more interdisciplinary and inter-sectorial R&D collaborations, and giving procurement at CERN the opportunity of reaching a wider selection of high-tech companies [L1]. CS concepts and techniques have been used for building the Technology Innovation Monitor (TIM) of the EC Joint Research Centre (JRC) [L2] and visualising compatibility and dependency relationships in software and metadata of the LHCb experiment at CERN [L1]. A collaboration with Wigner MTA and Budapest University of Technology and Economics (BME), Hungary is now in place to explore the use of CS in the areas of: (i) pharmacoinformatics as a supporting tool for knowledge-based drug discovery using analytics over linked open data for the life sciences [L3]; (ii) IT-analytics as a tool to enhance large-scale visual analysis of performance metrics of algorithmic and infrastructure components [L4]; (iii) neuroscience to assist in understanding the structural and functional organisation of the brain [L5]; and (iv) in social sciences as a research tool to study a database of European and international higher education institutes with the goal of developing new metrics to describe their performance (BRRG) [L6].

From the experience acquired with the prototypes, we plan to further develop the CS platform to provide optimized visual perception in order to enhance cognitive insights regardless of the size of the networks. These new developments will be tested at CERN and at the Wigner GPU laboratory.

Links:
[L1] collspotting.web.cern.ch/
[L2] timanalytics.eu/
[L3] bioinformatics.mit.bme.hu/UKBNetworks/
[L4] inf.mit.bme.hu/en/research/directions
[L5] kwz.me/hT6
[L6] kwz.me/hT7

Please contact:

Jean-Marie Le Goff, CERN, Switzerland
+41 22 767 6559, Jean-Marie.Le.Goff@cern.ch

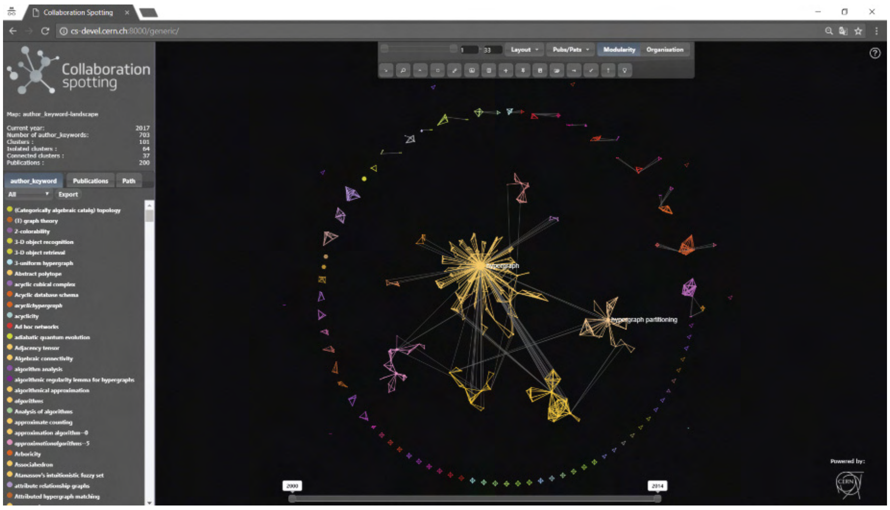


Figure 2: View of an analysis output depicting a network of publications from the keyword perspective. Vertices in italics represent keywords merged together and the others single-keywords. Coloured clusters highlight groups of keywords that are found more often together in publications.

Distributional Correspondence Indexing for Cross-Lingual and Cross-Domain Sentiment Classification

by Alejandro Moreo Fernández, Andrea Esuli and Fabrizio Sebastiani

Researchers from ISTI-CNR, Pisa (in a joint effort with the Qatar Computing Research Institute), have developed a transfer learning method that allows cross-domain and cross-lingual sentiment classification to be performed accurately and efficiently. This means sentiment classification efforts can leverage training data originally developed for performing sentiment classification on other domains and/or in other languages.

Sentiment Classification (SC) is the task of classifying opinion-laden documents in terms of the sentiment (positive, negative, or neutral) they express towards a given entity (e.g., a product, a policy, a political candidate). Determining the user's stance towards such an entity is of the utmost importance for market research, customer relationship management, the social sciences, and political science, and several automated methods have been proposed for this purpose. SC is usually accomplished via supervised learning, whereby a sentiment classifier is trained on manually labelled data. However, when SC needs to deal with a completely new context, it is likely that the amount of manually labelled, opinion-laden documents available to be used as training data, is scarce or even non-existent. Transfer Learning (TL) is a class of context adaptation methods that focus on alleviating this problem by allowing the learning algorithm to train a classifier for a "target" context by leveraging manually labelled examples available from a different, but related, "source" context.

In this work the authors propose the Distributional Correspondence Indexing (DCI) method for transfer learning in sentiment classification. DCI is inspired by the "Distributional Hypothesis", a well-know principle in linguistics that states that the meaning of a term is somehow determined by its distribution in text, and by the terms it tends to co-occur with (a famous motto that describes this assumption is "You shall know a word by the company it keeps"). DCI mines, from unlabelled datasets and in an unsupervised fashion, the distributional correspondences between each term and a small set of "pivot" terms, and is based on the further hypothesis that these correspondences are approximately invariant across the source context and target context for terms that play equivalent roles in the two contexts. DCI derives term representations in a vector space common to both contexts, where each dimension (or "feature") reflects its distributional correspondence to a pivot term. Term correspondence is quantified by means of a distributional correspondence function (DCF); in this work the authors propose and experiment with a number of efficient DCFs that are motivated by the distributional hypothesis. An advantage of DCI is that it projects each term into a low-dimensional space (about 100 dimensions) of highly predictive concepts (the pivot terms), while other competing methods (such as Explicit Semantic Analysis) need to work with much higher-dimensional vector spaces.

The authors have applied this technique to sentiment classification across domains (e.g., book reviews vs DVD reviews) and across languages (e.g., reviews in English vs reviews in German), either in isolation (cross-domain SC, or cross-lingual SC) or – for the first time in the literature – in combination (cross-domain cross-lingual SC, see Figure 1). Experiments run by the authors (on "Books", "DVDs", "Electronics", "Kitchen Appliances" as the domains, and on English, German, French, and Japanese as the languages) have shown that DCI obtains better sentiment classification accuracy than current state-of-the-art techniques (such as Structural Correspondence Learning, Spectral Feature Alignment, or Stacked Denoising Autoencoder) for cross-lingual and cross-domain sentiment classification. DCI also brings about a significantly reduced computational cost (it requires modest computational resources, which is an indication that it can scale well to huge collections), and requires a smaller amount of human intervention than competing approaches in order to create the pivot set.

Link: http://jair.org/papers/paper4762.html

Please contact:
Fabrizio Sebastiani, ISTI-CNR, Italy
+39 050 6212892, fabrizio.sebastiani@isti.cnr.it

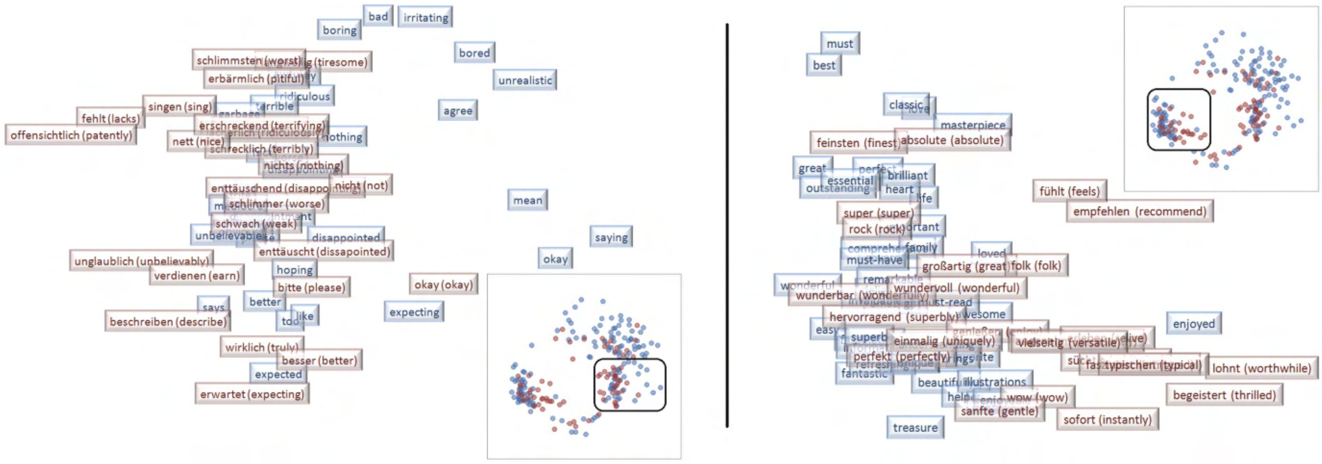


Figure 1: Cross-lingual (English-German) and cross-domain (book-music) alignment of word embeddings through DCI. The left part is a cluster of negative-polarity words, the right part of positive-polarity ones.

Measuring Bias in Online Information

by Evaggelia Pitoura (University of Ioannina), Irini Fundulaki (FORTH) and Serge Abiteboul (Inria & ENS Cachan)

Bias in online information has recently become a pressing issue, with search engines, social networks and recommendation services being accused of exhibiting some form of bias. Here, we make the case for a systematic approach towards measuring bias. To this end, we outline the necessary components for realising a system for measuring bias in online information, and we highlight the related research challenges.

We live in an information age where the majority of our diverse information needs are satisfied online by search engines, social networks, e-shops, and other online information providers (OIPs). For every request we submit, a combination of sophisticated algorithms return the most relevant results tailored to our profile. These results play an important role in guiding our decisions (e.g., what should I buy), in shaping our opinions (e.g., who should I vote for), and in general in our view of the world.

Undoubtedly, although the various OIPs have helped us in managing and exploiting the available information, they have limited our information seeking abilities, rendering us overly dependent on them. We have come to accept such results as the "de facto" truth, immersed in the echo chambers and filter bubbles created by personalisation, rarely wondering whether the returned results represent a full range of viewpoints. Further, there are increasingly frequent reports of OIPs exhibiting some form of bias. In the recent US presidential elections, Google was accused of being biased against Donald Trump [L1] and Facebook of contributing to the post-truth politics [L2]. Google search has been accused of being sexist or racist when returning images for queries such as "nurse" [L3]. Similar accusations have been made for Flickr, Airbnb and LinkedIn.

The problem has attracted some attention in the data management community [1, 2], and is also considered a high-priority problem for machine learning algorithms [3] and AI [L4]. Here we focus on the very first step, that of defining and measuring bias, by proposing a systematic approach for addressing the problem of bias in online information. According to the Oxford English Dictionary [L5], bias is "an inclination or prejudice for or against one person or group, especially in a way considered to be unfair", and as "a concentration on or interest in one particular area or subject".

When it comes to bias in OIPs, we make the distinction between subject bias and object bias. Subject bias refers to bias towards the users that receive a result, and it appears when different users receive different content based on user attributes that should be protected, such as gender, race, ethnicity, or religion. On the other hand, object bias refers to biases in the content of the results for a topic (e.g., USA elections), that appears when a differentiating aspect of the topic (e.g., political party) is disproportionately represented in the results.

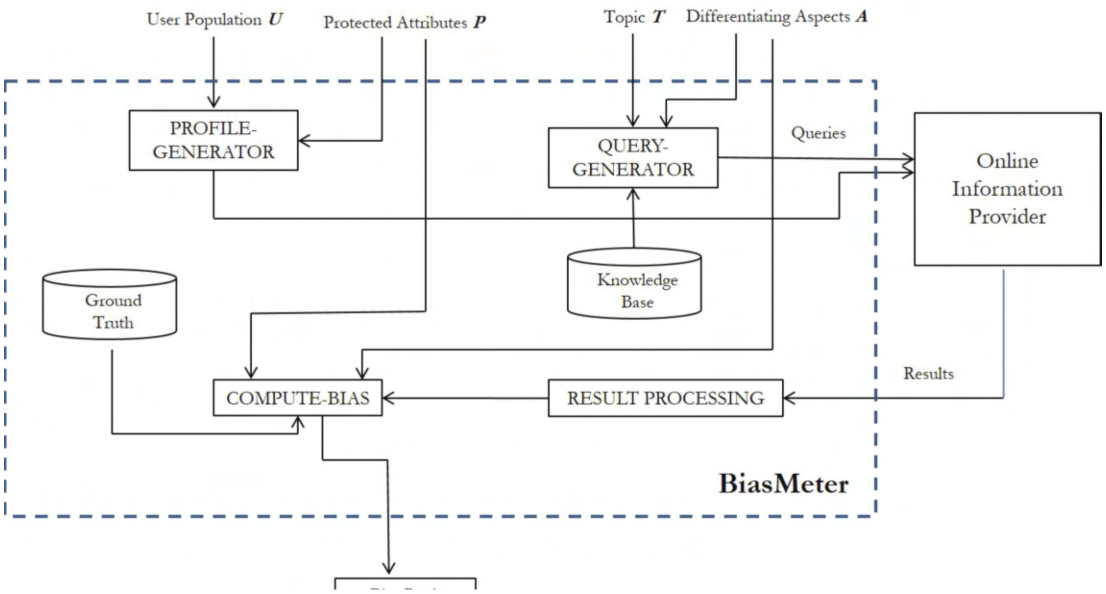


Figure 1: System Components of BiasMeter.

Let us now describe the architecture and the basic components of the proposed system BiasMeter. We treat the OIP as a black-box, accessed only through its provided interface, e.g., search queries. BiasMeter takes as input: (i) the user population U for which we want to measure the (subject) bias; (ii) the set P of the protected attributes of U; (iii) the topic T for which we want to test the (object) bias; and (iv) the set A of the differentiating aspects of the topic T. We assume that the protected attributes P of the user population and the differentiating aspects A of the topic are given as input (a more daunting task would be to infer these attributes).

Given the topic T and the differentiating aspects A, the goal of the Query Generator (supported by a knowledge base) is to produce an appropriate set of queries to be fed to OIP that best represent the topic and its aspects. The Profile Generator takes as input the user population U and the set of protected attributes P, and produces a set of user profiles appropriate for testing whether the OIP discriminates over users in U based on the protected attributes in P (e.g., gender). The Result Processing component takes as input the results from the OIP and applies machine learning and data mining algorithms such as topic modelling and opinion mining to determine the value of the differentiating aspects (e.g., if a result takes a positive stand). Central to the system is the Ground Truth module, but obtaining the ground truth is hard in practice. Finally, the Compute Bias component calculates the bias of the OIP based on the subject and object bias metrics and the ground-truth.

Since bias is multifaceted, it might be difficult to quantify it. In [L6] we propose some subject and object bias metrics. Further, obtaining the ground truth and the user population are some of the most formidable tasks in measuring bias, since it is difficult to find objective evaluators and generate large samples of user accounts for the different protected attributes. Also there are many engineering and technical challenges for the query generation and result processing components that involve knowledge representation, data integration, entity detection and resolution, sentiment detection, etc. Finally, it might be in the interest of governments to create legislation that provides access to sufficient data for measuring bias, since access to the internals of OIPs is not provided.

Links:
[L1] https://kwz.me/hLc
[L2] https://kwz.me/hLy
[L3] https://kwz.me/hLH
[L4] https://futureoflife.org/ai-principles/
[L5] https://en.oxforddictionaries.com/definition/bias
[L6] http://arxiv.org/abs/1704.05730

References:

[1] J. Stoyanovich, S. Abiteboul, and G. Miklau. Data, responsibly: Fairness, neutrality and transparency in data analysis. In EDBT, 2016.

[2] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, I. Shibpur, I. K. P. Gummadi, and K. Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In CSCW, 2017.

[3] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In WWW, 2017.

Please contact:

Irini Fundulaki, Foundation for Research and Technology – Hellas, Institute of Computer Science, Greece,
fundul@ics.forth.gr

The Approximate Average Common Submatrix for Computing the Image Similarity

by Alessia Amelio (Univ. of Calabria)

The similarity between images can be computed using a new method that compares image patches where a portion of pixels is omitted at regular intervals. The method is accurate and reduces execution time relative to conventional methods.

To date, researchers have not quite solved the problem of automatically computing the similarity between two images. This is mainly due to the difficulty of filling in the gap between the human visual similarity and the similarity which is captured by the machine. In fact, it requires two important objectives to be fullfilled: (i) to find a reliable and accurate image descriptor that can capture the most important image characteristics, and (ii) to use a robust measure to evaluate similarity between the two images according to their descriptors. Usually a trade-off is needed between these two objectives and the execution time on the machine. An important challenge, therefore, is to achieve computation of an accurate descriptor and a robust measure whilst reducing execution time.

In this paper, I present a new measure for computing the similarity between two images which is based on the comparison of image patches where a portion of pixels is omitted at regular intervals. This measure is called Approximate Average Common Submatrix (A-ACSM) [1] and it is an extension of the Average Common Submatrix (ACSM) measure [2]. The advantage of this similarity measure is that it does not need to extract complex descriptors from the images to be used for the comparison. On the contrary, an image is considered as a matrix, and the similarity is evaluated by measuring the average area of the largest sub-matrices which the two images have in common. The principle underlying this evaluation is that two images can be considered as similar if they share large patches representing image patterns. Two patches are considered as identical if they match in a portion of the pixels which are extracted at regular intervals along the rows and columns of the patches. Hence, the measure is an easy match between a portion of the pixels. This concept introduces an approximation, which is based on the “naive” consideration that two images do not need to exactly match in the intensity of every pixel to be considered as similar. This approximation makes the similarity measure more robust to noise, i.e., small variations in the pixels’ intensity due to errors in image generation, and considerably reduces its execution time when it is applied on large images, because a portion of the pixels does not need to be checked. Figure 1 shows a sample of match between two image patches with an interval of two along the columns and one along the rows of the patches.

Figure 2 depicts the algorithm for computing the A-ACSM similarity measure.

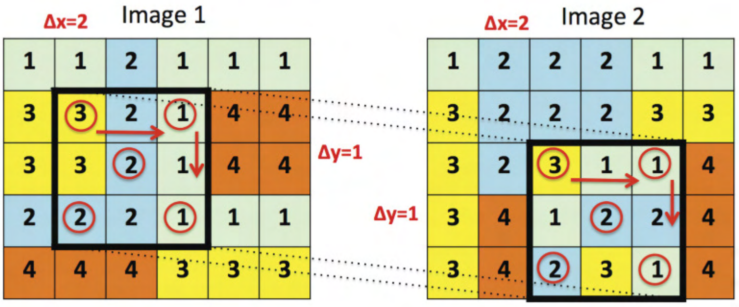


Figure 1: A demonstration sample of match between two patches belonging to images 1 and 2. Each image has four colours which are also numbered from 1 to 4. An interval of Δx=2 and Δy=1 is set respectively along the columns and rows of the patches. Accordingly, the match is only verified between the elements in the red circles. The elements are selected as in a chessboard. In this case, the two image patches perfectly match because all the elements in the red circles correspond to one another.

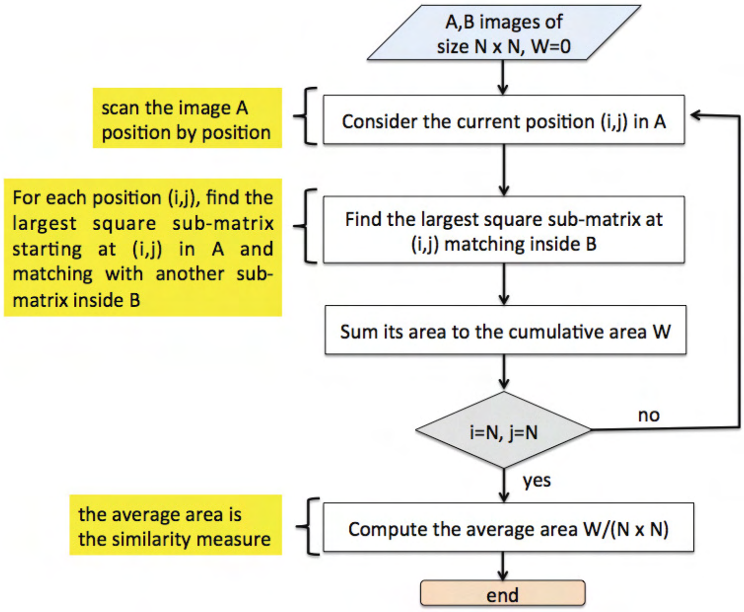

A,B images of size N x N, W=0
scan the image A position by position
Consider the current position (i,j) in A
For each position (i,j), find the largest square sub-matrix starting at (i,j) in A and matching with another sub-matrix inside B
Find the largest square sub-matrix at (i,j) matching inside B
Sum its area to the cumulative area W
i=N, j=N
no
yes
the average area is the similarity measure
Compute the average area W/(N x N)
end

Figure 2: Flowchart of the A-ACSM algorithm.

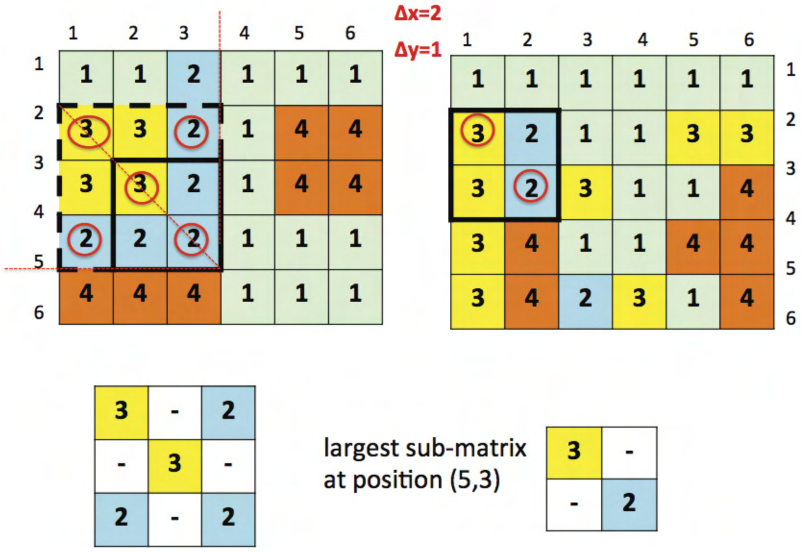


Figure 3: The largest square sub-matrix at position (5,3) of the first image. All the sub-matrices of different size along the main diagonal at (5,3) are considered to match inside the second image. In this case, the sub-matrix of size 3 has no match inside the second image. Hence, the sub-matrix of size 2 is considered. Because it has a correspondence at position (3,2) in the second image, it is the largest square sub-matrix at position (5,3).

Figure 3 shows how to find the largest square sub-matrix at a sample position (5,3) of the first image.

The A-ACSM similarity measure, as well as its corresponding dissimilarity measure, has been extensively tested on benchmark image databases and compared with the ACSM measure and other well-known measures in terms of accuracy and execution time. Results demonstrated that A-ACSM outperformed its competitors, obtaining higher accuracy in a lower execution time.

The project of the average common submatrix measures is in its early stage at the Georgia Institute of Technology, USA, and is currently in progress at DIMES University of Calabria, Italy. Future work will extend the submatrix similarity measures with new features and will evaluate the similarity on different types of data, including documents, sensor data, and satellite images [3]. The recent developments in this direction are in collaboration with the Technical Faculty in Bor, University of Belgrade, Serbia.

References:

[1] A. Amelio: "Approximate Matching in ACSM dissimilarity measure", Procedia Computer Science 96: 1479-1488, 2016.

[2] A. Amelio, C. Pizzuti: "A patch-based measure for image dissimilarity", Neurocomputing 171:362-378, 2016.

[3] A. Amelio, D. Brodić: The ε-Average Common Submatrix: Approximate Searching in a Restricted Neighborhood, IWCIA: 7-11, (short comm.) arXiv:1706.06026, 2017.

Please contact:

Alessia Amelio, DIMES University of Calabria, Italy
aamelio@dimes.unical.it

My-AHA: Stay Active, Detect Risks Early, Prevent Frailty!

by Nadine Sturm, Doris M. Bleier and Gerhard Chroust (Johanniter Österreich Ausbildung und Forschung gemeinnützige GmbH)

Aging of the population is a major concern in the European Union. "AHA – Active and Healthy Aging" is a multi-facetted and systemic approach in order to compensate for the diminishing capabilities of seniors. The project my-AHA is based on reducing frailty through targeted, personalised ICT-based interventions.

With the demographic changes that are occurring, particularly within the European Union, we need to consider how best to support the ageing population. The broad concept of active and healthy ageing (AHA) was proposed by the World Health Organisation (WHO) as an answer. Most seniors will by necessity become actively and/or passively involved in activities related to AHA. The goal of AHA can be roughly described as trying to compensate for the diminished capabilities of seniors related to age. It is a highly interdisciplinary challenge which involves the following areas:

- Physiological/medical (maintaining physical health);
- Psychological (preserving mental health and a positive attitude to the world and seniors, avoiding the trauma of aging);
- Sociological (maintaining a good relationship between the individual and his/her immediate support environment, e.g., formal and informal caregivers, healthcare services and infrastructure);
- Technological (providing technological support on the physical and mental level, especially by ICT, sensors, activators, etc.);
- Organisational (providing logistic support and infrastructure on various levels, including architectural considerations);
- Economic (defraying cost of treatment, supplying equipment, etc.).

The main goal of the European Project "my-AHA" (My Active and Healthy Aging, European Union – Horizon-2020 Project No 689592, 2016-2019 [L1] is to reduce the risk of frailty by improving physical activity and cognitive function, psychological state, social resources, nutrition and sleep, thereby contributing to overall well-being. An ICT-based platform, including a smartphone system, will detect defined personalised risks for frailty, be they physiological, psychological or social, early and accurately via non-stigmatising embedded sensors and data readily available in the daily living environment of older adults.

When a risk is detected, my-AHA will provide personalised targeted ICT-based interventions with scientific evidence based efficacy, including vetted offerings from established providers of medical and AHA support. These interventions will follow an integrated approach to motivate users to participate in exercises, cognitively stimulating games and social networking to achieve long-term behavioural change, sustained by continued end user engagement with my-AHA. This approach will also increase a senior's competence with respect to his/her own frailties. It encourages long-term changes to a senior's behaviour and as a result leads to a longer, more active, more healthy and thus more enjoyable life.

Smartphone for interventions.

The project is coordinated by the University degli Studi di Torino, Italy, and has 14 other partners from: Austria (one), Australia (one), Germany (four), Italy (one), Japan (two), Netherlands (one), Portugal (one), South Korea (one), Spain (one), and United Kingdom (one).

The Johanniter Austria Research and Education (Johanniter Österreich Ausbildung und Forschung gemeinnützige GmbH) [L2] provide considerable expertise in methodologies of empirical social science, especially with respect to user involvement, honouring ethical considerations and applying technological foresight. This input will be invaluable when it comes to developing targeted interventions and tests and when introducing new technologies and identifying the most effective methods to prevent frailty. Additionally, the Johanniter are setting up a Europe-wide testing environment with standardised parameters to support the project.

Links:
[L1] www.activeageing.unito.it/home
[L2] https://www.johanniter.at/dienste/forschung/

Please contact:

Georg Aumayr
Johanniter Österreich – Ausbildung und Forschung gemeinnützige GmbH, Austria
+43 1 470 7030 2222
georg.aumayr@johanniter.at

A New Vision of the Cruise Ship Cabin

by Erina Ferro (ISTI-CNR)

The National Research Council of Italy (CNR) is involved in the E-Cabin project, which aims to create a futuristic and intelligent cabin in order both to improve the well-being and satisfaction of passengers and to minimize the on-board waste by using advanced consumption control.

The E-Cabin project focusses on the cabins of cruise ships, with the aim of creating a set of advanced technology solutions to enhance the travel experience both inside the cabin and throughout the entire ship, providing the ship owner with an additional monitoring system for each cabin. A cruise ship is now a concentration of technologies, equipment, communication and security systems that is hard to find in terrestrial spaces. However, the existing set of technologies is not enough to improve the experience of individual passengers, who require both a continuous connection with the outside world and personalised tools to best enjoy the opportunities provided by the ship system. At the same time, new technological solutions can also be proposed to the shipbuilding company, thus increasing the control and safety systems.

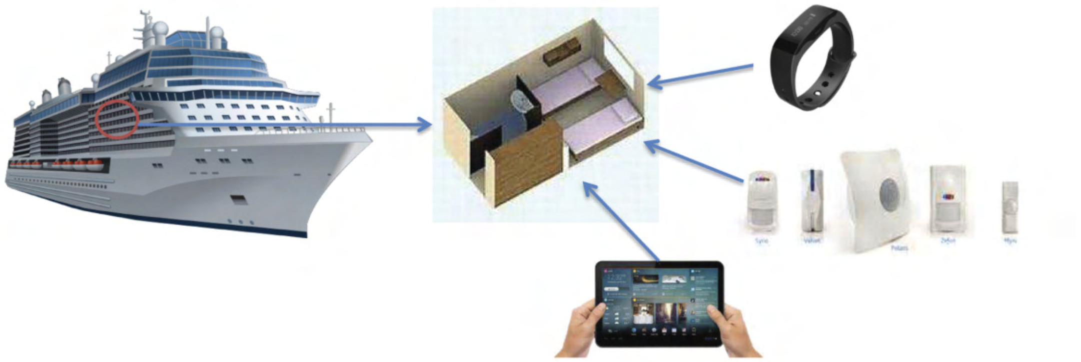


Figure 1: The E-Cabin.

In order to create a futuristic and intelligent cabin, E-Cabin has two primary objectives, one for the passenger and one for the shipbuilding company. In the first case, the aim is to improve a passenger's well-being and satisfaction by making his presence on board more enjoyable thanks to a set of innovative services. In the second case, the goal is to minimise the on-board waste thanks to a precise consumption control (for example light, heating/cooling, etc.) by means of a set of solutions that include:

1. To carry out a cab monitoring system in order to understand consumption, to plan the maintenance work and to dynamically manage the ship's resources. This will be obtained via a set of sensors installed in the cab, whose battery life will be possibly extended through an energy harvesting system.
2. To understand and take advantage of the correlations and the interactions between the indexes that govern feelings of comfort so as to maximise the overall comfort of passengers. This information will be collected from heterogeneous devices: sensors and actuators installed in the cabin and personal user devices (smartphones, smartwatches and/or wearable sensors) in order to detect the correct correlation between the personal feeling of well-being and the environmental data (intensity of lights, noise, temperature, humidity, etc.).
3. To realise a set of applications that learn the habits of passengers, thus predicting their needs, increasing their opportunities to socialise, sharing contents through mobile social networking applications, and enriching their participation in the "ship world" by taking advantage of augmented reality information relevant to the cruise.

The global e-cabin system is depicted in Fig. 1. The real novelty of this approach is that all solutions will be integrated into a single development platform for the various applications, which will be primarily related to the quality of life in the cabin, but which will also interact with applications related to other activities carried out on the ship. Vertical solutions, which do not communicate each other (typical of proprietary business solutions), will not be adopted. This way, all collected data will be related, thus generating additional knowledge. Just as well-being is the result of a combination of different sensations, physical factors, and environmental factors, applications developed in E-Cabin and applied technologies must also be able to combine and share data and information in order to customise solutions.

On-board energy consumption will be monitored and automatically reduced by implementing waste reduction policies based on strict waste control; this aspect is of particular interest to the ship owner. All developed components will be miniaturised and integrated into the entire E-Cabin system. The E-Cabin prototype can be fully engineered and integrated in time with new features and applications that can be matched with evolving needs and technology.

The experimentation will take place at Fincantieri in Trieste (Italy). This project involves the ISTI-CNR, IIT-CNR, ISTEC-CNR, ITIA-CNR, IEIIT-CNR Institutes and the University of Trieste.

Please contact:
Erina Ferro, ISTI-CNR, Italy
erina.ferro@isti.cnr.it

Trend Analysis of Underground Marketplaces

by Klaus Kieseberg, Peter Kieseberg and Edgar Weippl: (SBA Research)

Underground marketplaces, which represent one of the most prominent examples for criminal activities in the Darknet, form their own economic ecosystems, often connected to cyber-attacks against critical infrastructures. Retrieving first-hand information on emerging trends in these ecosystems is thus of vital importance for law enforcement, as well as critical infrastructure protection.

The term “Darknet” generally describes “overlay networks” that are only accessible to a few exclusive users, but it is often used in order to describe parts of the internet that are sealed off like underground marketplaces, or closed peer-to-peer networks. These networks, and their potential links to criminal and terrorist activities, have recently gained public attention, which has highlighted the need for an efficient analysis of Darknets and similar networks.

We intend to study how these underground forums operate as a means for unobserved communication between like-minded individuals as well as a tool for the propagation of political propaganda and recruitment. We also focus heavily on Darknets used for trading illegal goods and especially services that could be used to attack government institutions and undermine national security. For example, in some of these networks it is possible to buy the services of bot networks for launching “Distributed Denial of Service” (DDoS) attacks against sensitive infrastructures like power distribution networks, as well as physical goods like drugs and arms. An efficient analysis of these underground marketplaces is therefore essential for the prevention of terrorist attacks and to stem the proliferation of digital weapons.

In the course of our research, which notably focused on trend analysis in underground marketplaces, the following three key issues emerged that require special attention:

1. Detection and analysis of data sources: In order to get a good database for subsequent analysis, a detailed source analysis and source detection regarding propaganda and illegal services, as well as an assessment of sources regarding their relevance concerning national security is required. This also contains means for the undetected automation of data collection, in order to get undistorted information and to not compromise the information gathering process.
2. Privacy preserving analysis of data: Due to the strict privacy requirements for data processing put forth by the “General Data Protection Regulation” (GDPR) [1] new techniques in privacy preserving machine learning have to be invented. In addition, techniques for monitoring access to the crawled data as well as methods for manipulation detection need to be developed.
3. Studying the mode of operation of underground marketplaces: This includes the mechanisms of establishing first contact, pricing, payment and the transfer of goods, particularly goods that would require some sort of contact in the offline world [2]. A special interest for securing critical infrastructures lies in the analysis of trends, rather than individual behaviour, as this is more interesting from a strategic point of view and far less problematic with respect to sensitive information.

In addition to the technical problems, the analysis of this type of information also opens up many important legal questions. Especially the new rules and regulations introduced by the GDPR (General Data Protection Regulation) and the national counterparts play a significant part, since they aim to ensure transparency of data processing procedures concerning personal information, as well as the possibility to delete data from data processing applications.

When it comes to analysing underground marketplaces regarding the trade of illegal goods or services, it is important to develop techniques that can be automated to a certain degree but still integrate a human component to detect and evaluate criminal activities. Furthermore, many underground forums have detection mechanisms in place to identify automated information gathering, as well as users with strange access behaviour that hints at them belonging to law enforcement [2]. Often, manual intervention in the form of messages with questions, as well as the analysis of access patterns are utilised by the forum owners. Thus, methods must be developed that mask the information gathering process and emulate user behaviour. To maximise the level of automation achievable, it is convenient to focus on the detection and identification of trends instead of investigating every individual case in isolation.

A vital aspect of this research work is to ensure anonymity and to secure the privacy of innocent people. We therefore strive to develop techniques and methods that allow for an efficient analysis of underground marketplaces while providing ample protection to people not involved in illegal activities. The collection of data will be a selective process adhering to the “data minimisation principle” introduced by the GDPR. Instead of collecting and analysing all available data, an intelligent collection process is required, where data is first evaluated regarding its importance and selected accordingly. This also includes metadata, which is of low importance for trend analysis itself, but can contain important information, as well as useful links between individual information particles. This not only increases the performance of the data collection process, but also minimises the amount of data collected while providing additional insights into the overall ecosystem of underground markets.

Another important question is how privacy protection mechanisms can affect conventional machine learning techniques. Methods used to protect sensitive information like anonymization of data via k-anonymity, as well as deleting individual records from data sets, can alter the information derived from the data analysis. While these are key factors in choosing the right protection method, as well as a suitable security factor, these effects have not yet been studied thoroughly enough [3]. In the course of our work, we examine the effects of different protective measures on machine learning techniques and develop methods to mitigate them.

Based on these results, we will develop means for controlling the introduced error by: (i) providing upper bounds for the effects of anonymization and deletion, as well as by (ii) developing new methods for analysing specific forms of anonymized sensitive information that introduce less distortion into the final result.

In conclusion, the analysis of underground marketplaces and other areas typically considered to constitute the Darknet requires additional research effort in order to deliver the information required for analysing trends and the workings of their market mechanisms. This led to the development of the “Darknet Analysis”-project [L1], which is currently tackling the research questions outlined above. Furthermore, the results of our research will make an important contribution to the topic of privacy protection in law enforcement as a whole and thus have high re-use value for the involved governmental stakeholders.

Link:
[L1] http://www.kiras.at/gefoerderte-projekte/

References:

[1] Regulation (EU) 2016/679 Of The European Parliament And Of The Council Of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[2] H. Fallmann, G. Wondracek, C. Platzer. “Covertly Probing Underground Economy Marketplaces”, DIMVA, Vol. 10, 2010.

[3] B. Malle, P. Kieseberg, E. R. Weippl, A. Holzinger: “The right to be forgotten: towards machine learning on perturbed knowledge bases”, in International Conference on Availability, Reliability, and Security (pp. 251-266), Springer, 2016.

Please contact:

Peter Kieseberg
SBA Research, Vienna, Austria
pkieseberg@sba-research.org

RDF Visualiser: A Tool for Displaying and Browsing High Density of RDF Data

by Nikos Minadakis, Kostas Petrakis, Korina Doerr and Martin Doerr (FORTH)

RDF Visualiser is a generic browsing mechanism that gives the user a flexible, highly configurable, detailed overview of an RDF dataset / database, designed and developed to overcome the drawbacks of the existing RDF data visualisation methods/tools. RDFV is currently used in EU research projects by the 3M editor of the X3ML suite of tools and has been tested with large datasets from the British Museum and the American Art Collaborative project.

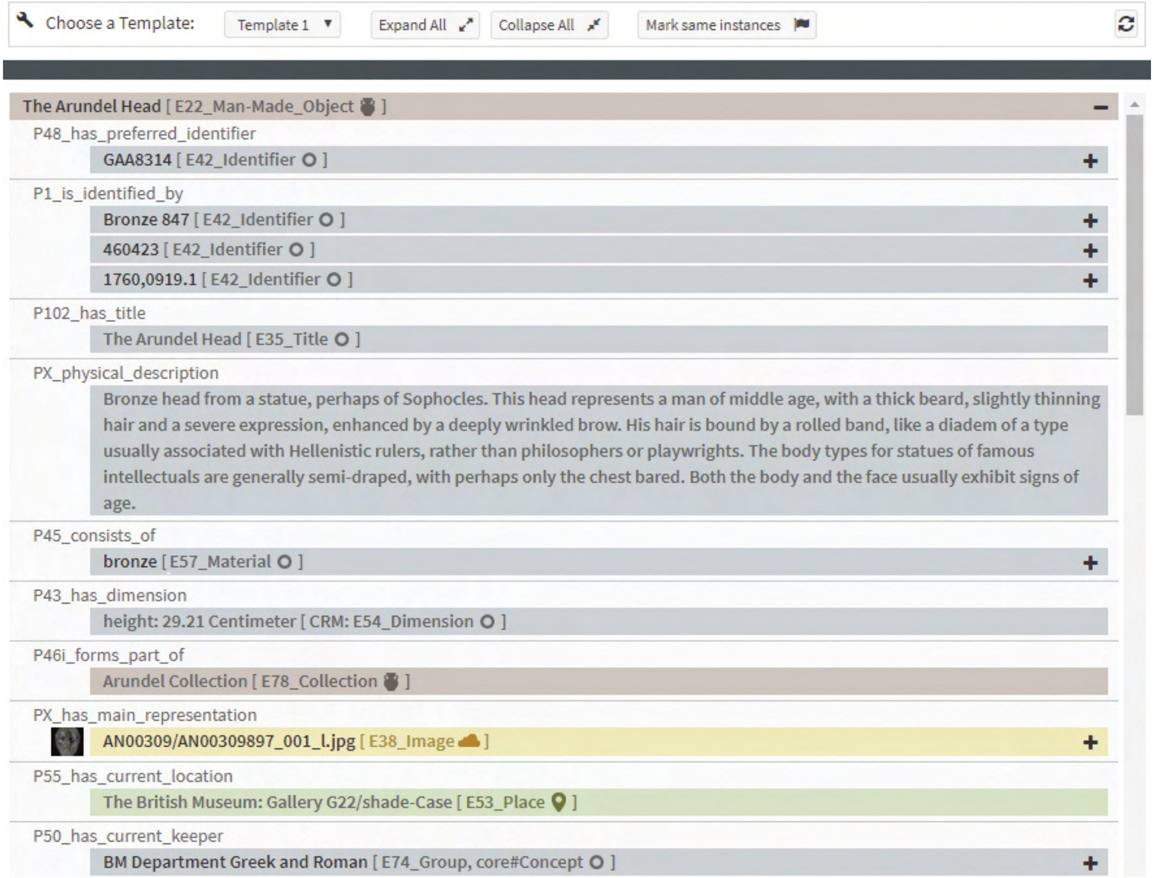
RDF is widely used for data integration, transformation and aggregation of heterogeneous sources and applications of mappings between the source schemata. Although a great deal of emphasis has been placed on the validation of the produced RDF structure and format, the efficient visualisation of the constructed database contents that enables semantic validation by domain experts has largely been ignored and is achieved either by manual inspection of multiple files, by formulation and execution of complex SPARQL queries, or by custom user interfaces that work only with a particular RDF schema without intervention by a programmer.

The basic principles that an efficient and user-friendly RDF data visualisation tool should be able to live up to are:
1. The ability to display data of any schema and RDF format.
2. The ability to display all nodes of any class/instance.
3. The application of configuration rules to improve the layout or presentation for known classes and properties (e.g., hide URIs that are meaningless for the user).
4. The display of a high density of information in one screen (which is not possible in solutions based on “object templates”).

Although some visualisation solutions already exist and have been applied to successful projects [L1],[1], the existing approaches fail to fulfil the complete set of principles enumerated above. Specifically, approaches that display all the nodes of any class/schema and support any schema and format (principles 1 and 2) fail to display a high density of information in one screen and those that succeed with regards to the latter fail in relation to the rest of the principles.

In order to meet the requirements laid out by the complete set of principles, our team designed and implemented the RDF Visualiser (RDFV), a generic browsing mechanism that gives the user a flexible, highly configurable, detailed overview of an RDF dataset / database.

RDFV presents RDF data as an indented list to handle the density and depth of information (principle 4) starting from a specified RDF resource (URI). In order to achieve this, all incoming links are inverted to display all the nodes of every



The Arundel Head [E22_Man-Made_Object]

P48_has_preferred_identifier
GAA8314 [E42_Identifier ○]

P1_is_identified_by
Bronze 847 [E42_Identifier ○]
460423 [E42_Identifier ○]
1760,0919.1 [E42_Identifier ○]

P102_has_title
The Arundel Head [E35_Title ○]

PX_physical_description
Bronze head from a statue, perhaps of Sophocles. This head represents a man of middle age, with a thick beard, slightly thinning hair and a severe expression, enhanced by a deeply wrinkled brow. His hair is bound by a rolled band, like a diadem of a type usually associated with Hellenistic rulers, rather than philosophers or playwrights. The body types for statues of famous intellectuals are generally semi-draped, with perhaps only the chest bared. Both the body and the face usually exhibit signs of age.

P45_consists_of
bronze [E57_Material ○]

P43_has_dimension
height: 29.21 Centimeter [CRM: E54_Dimension ○]

P46i_forms_part_of
Arundel Collection [E78_Collection]

PX_has_main_representation
AN00309/AN00309897_001_l.jpg [E38_Image]

P55_has_current_location
The British Museum: Gallery G22/shade-Case [E53_Place]

P50_has_current_keeper
BM Department Greek and Roman [E74_Group, core#Concept ○]

Figure 1: RDFV user interface.

class/instance (principle 2) in a schema agnostic way (principle 1). Users are able to configure the display of schema-dependent information according to their preferences (principle 4). By editing an xml file, users are also able to define priority based rule chains that are used to define schema-dependent style and order of properties that are inherited to subclasses and subproperties (principle 4). For anything not covered by a rule, default options are applied based on our experience and best practices. The user interface of the tool has been designed in cooperation with potential users, with a focus on usability and readability (Figure 1).

Moreover, rich functions are provided to the user to control the display of data items, such as identifying same instances, expanding collapsing big texts or big sets of results, selecting the maximal depth of information, displaying images and image galleries, removing prefixes, selecting non displayed URIs, retrieving the path of the sub-graph for a specific node etc.

RDFV supports browsing of content in triple stores (tested successfully on Virtuoso and BlazeGraph) and in local files of every format. It has been integrated in the 3M interface of the X3ML suite of tools for data mapping and transformation [2]. Added into the 3M interface it adds an important validation tool for data mapped and transformed by domain experts who wish to check and correct the resulting outputs, enabling an iterative and collaborative evaluation of the resultant RDF. RDFV is currently exploited by a number of EU research projects, and has been tested with large RDF datasets from the British Museum and from the American Art Collaborative project.

In the future RDFV will support geospatial display functions on maps along with a new refined set of the existing functions with a focus on configurability and personalisation.

Our team will continue to work on the development, maintenance and user support of RDFV. The RDFV components, which consist of API microservices and a web application, are also available as independent blocks to assist developers with building their own applications through GitHub by September 2017.

Link:
[L1] http://www.researchspace.org/

References:

[1] N. Minadakis, et al.: "LifeWatch Greece Data-Services: On Supporting Metadata and Semantics Integration for the Biodiversity Domain", in 13th International Congress on the Zoogeography and Ecology of Greece and Adjacent Regions (ICZEGAR'2015), Heraklion, Crete, October 2015.

[2] Y. Marketakis, et al.: "X3ML Mapping Framework for Information Integration in Cultural Heritage and beyond", International Journal on Digital Libraries, pp 1-19, Springer, DOI 10.1007/s00799-016-0179-1.

Please contact:

Minadakis Nikos, FORTH, Greece
minadakn@ics.forth.gr

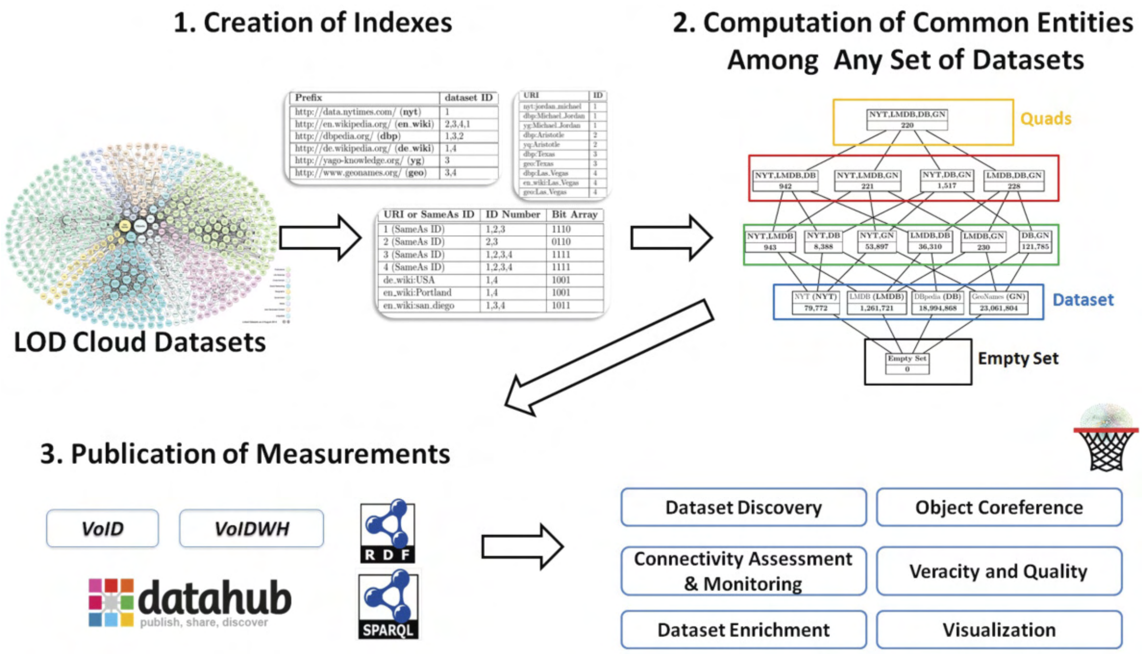
Services for Large Scale Semantic Integration of Data

by Michalis Mountantonakis and Yannis Tzitzikas (ICS-FORTH)

LODsyndesis is a new suite of services that helps the user to exploit the linked data cloud.

In recent years, there has been an international trend towards publishing open data and an attempt to comply with standards and good practices that make it easier to find, reuse and exploit open data. Linked Data is one such way of publishing structured data and thousands of such datasets have already been published from various domains.

However, the semantic integration of data from these datasets at a large (global) scale has not yet been achieved, and this is perhaps one of the biggest challenges of computing today. As an example, suppose we would like to find and examine all digitally available data about Aristotle in the world of Linked Data. Even if one starts from DBpedia (the database derived by analysing Wikipedia), specifically from the URI “http://dbpedia.org/resource/Aristotle” it is not possible to retrieve all the available data because we should first find ALL equivalent URIs that are used to refer to Aristotle. In the world of Linked Data, equivalence is expressed with “owl:sameAs” relationships. However, since this relation is transitive, one should be aware of the contents of all LOD datasets (of which there are currently thousands) in order to compute the transitive closure of “owl:sameAs”, otherwise we would fail to find all equivalent URIs. Consequently, in order to find all URIs about Aristotle, which in turn would be the lever for retrieving all data about Aristotle, we have to



Prefix	dataset ID
http://data.nytimes.com/ (**nyt**)	1
http://en.wikipedia.org/ (**en_wiki**)	2,3,4,1
http://dbpedia.org/ (**dbp**)	1,3,2
http://de.wikipedia.org/ (**de_wiki**)	1,4
http://yago-knowledge.org/ (**yg**)	3
http://www.geonames.org/ (**geo**)	3,4

URI	ID
nyt:jordan_michael	1
dbp:Michael_Jordan	1
yg:Michael_Jordan	1
dbp:Aristotle	2
yq:Aristotle	2
dbp:Texas	3
geo:Texas	3
dbp:Las_Vegas	4
en_wiki:Las_Vegas	4
geo:Las_Vegas	4

URI or SameAs ID	ID Number	Bit Array
1 (SameAs ID)	1,2,3	1110
2 (SameAs ID)	2,3	0110
3 (SameAs ID)	1,2,3,4	1111
4 (SameAs ID)	1,2,3,4	1111
de_wiki:USA	1,4	1001
en_wiki:Portland	1,4	1001
en_wiki:san_diego	1,3,4	1011

Figure 1: The whole process of LODsyndesis

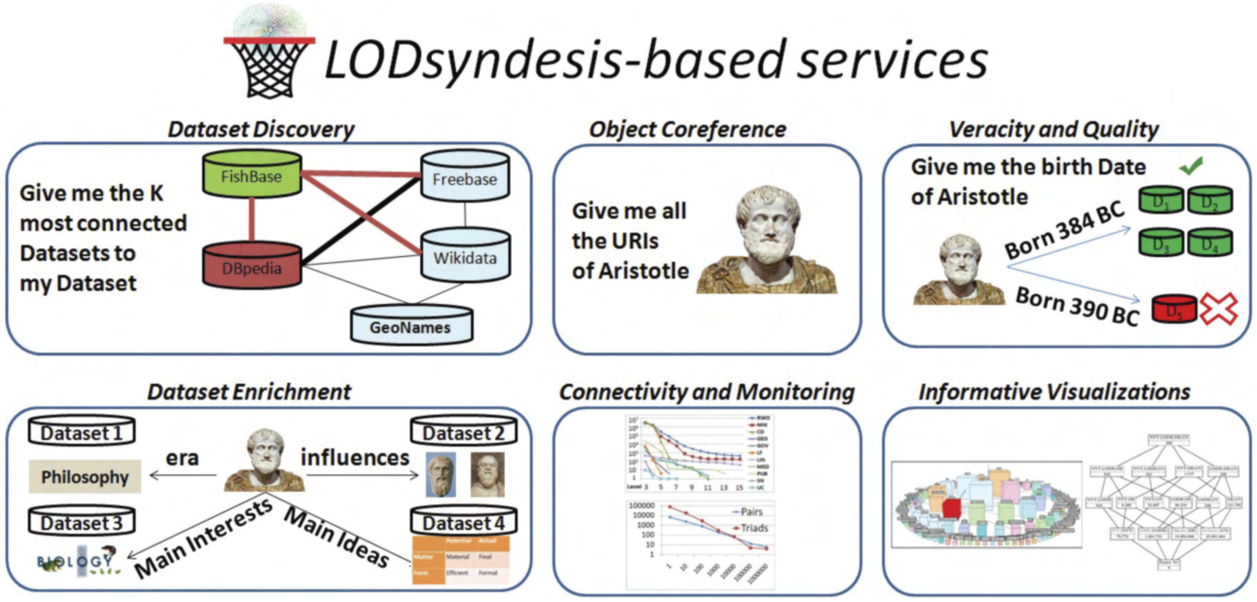


Figure 2: Services offered by LODsyndesis.

index and enrich numerous datasets, through cross-dataset inference.

.

The Information Systems Laboratory of the Institute of Computer Science of FORTH, designs and develops innovative indexes, algorithms and tools to assist the process of semantic integration of data at a large scale. This endeavour started two years ago, and the current suite of services and tools that have been developed are known as “LODsyndesis” [L1]. As shown in Figure 1, the distinctive characteristic of LODsyndesis is that it indexes the contents of all datasets in the Linked Open Data cloud [1]. It then exploits the contents of datasets to measure the commonalities among the datasets. The results of measurements are published on the web and can be used to perform several tasks. Indeed, “global scale” indexing can be used to achieve various outcomes, and an overview of the available LODsyndesis-based services for these tasks can be seen in Figure 2.

First, it is useful for dataset discovery, since it enables content-based dataset discovery services, e.g., it allows answering queries of the form: “find the K datasets that are more connected to a particular dataset”. Another task is object co-reference, i.e., how to obtain complete information about one particular entity (identified by a URI) or set of entities, including provenance information. LODSyndesis can also help with assessing the quality and veracity of data, since the collection of all information about an entity, and the cross-dataset inference that can be achieved, allows contradictions to be identified and provides information for data cleaning or for estimating and suggesting which data are probably correct or most accurate. Last but not least, these services can help to enrich datasets with more features that can obtain better predictions in machine learning tasks [2] and the related measurements can be exploited to provide more informative visualisation and monitoring services.

The realisation of the services of LODSyndesis is challenging because they presuppose knowledge of all datasets. Moreover, computing the transitive closure of all “owl:sameAs” relationships is challenging since most of the algorithms for doing this require a lot of memory. To tackle these problems LODsyndesis is based on innovative indexes and algorithms [1] appropriate for the needs of the desired services. The current version of LODsyndesis indexes 1.8 billion triples from 302 datasets, it contains measurements of the number of common entities among any combination of datasets (e.g., how many common entities exist in a triad of datasets), while the performed measurements are also available in DataHub [L2] for direct use. It is worth noting that currently only 38.2% of pairs of datasets are connected (i.e., they share at least one common entity) and only 2% of entities occur in three or more datasets. The aforementioned measurements reveal the sparsity of the current datasets of the LOD Cloud and justify the need for services for assisting integration at large scale. This method is efficient, the time for constructing the indexes and performing all these measurements being only 22 minutes using a cluster of 64 computers.

Over the next two years we plan to improve and extend this suite of services. Specifically, we plan to advance the data discovery services and to design services for estimating the veracity and trust of data.

This project is funded by FORTH and it has been supported by the European Union Horizon 2020 research and innovation programme under the BlueBRIDGE project (Grant agreement No 675680).

Links:
[L1] http://www.ics.forth.gr/isl/LODsyndesis/
[L2] https://datahub.io/dataset/connectivity-of-lod-datasets

References:

[1] M. Mountantonakis and Y. Tzitzikas, “On measuring the lattice of commonalities among several linked datasets,” Proceedings of the VLDB Endowment, vol. 9, no. 12, pp. 1101-1112, 2016.

[2] M. Mountantonakis and Y. Tzitzikas, “How Linked Data can aid Machine Learning-based Tasks,” in International Conference on Theory and Practice of Digital Libraries, 2017.

Please contact:

Yannis Tzitzikas
FORTH-ICS and University of Crete
+30 2810 391621
tzitzik@ics.forth.gr

On the Interplay between Physical and Digital World Accessibility

by Christophe Ponsard, Jean Vanderdonckt and Lyse Saintjean (CETIC)

Many means of navigating our physical world are now available electronically: maps, streets, schedules, points of interest, etc. While this “digital clone” is continually expanding both in completeness and accessibility, an interesting interplay scenario arises between physical and digital worlds that can benefit all of us, especially the mobility-impaired.

Physical world accessibility is about ensuring that features of physical places (e.g., shops, tourist attractions, offices) are designed to accommodate the (dis)abilities of people visiting them in order to ensure optimal access for all, including the estimated 15% of the population living with some kind of impairment. Our world is undergoing digitisation at an ever-increasing rate. Online maps are becoming so precise that the inner structure of buildings is often captured, tours are available, while locations are ranked by users, etc. As a consequence, many new opportunities have arisen for breaking accessibility barriers and better combining the physical and digital aspects of accessibility. For example, the use of the available online information can be complemented with physical measures to help assess the accessibility of buildings and modes of transport. Combining digital maps, crowd-sourced information, open data and expert assessments can help produce up-to-date accessibility data at a larger scale and reduced cost.

Physical accessibility requirements are now well defined and documented, such as in the ISO21542 standard. Assessing physical accessibility is not straightforward because it requires the identification of physical obstacles to accessibility relevant to specific kinds of impairment, e.g. a doorstep for a wheelchair, the presence or not of braille on lift buttons for blind people, the printing of light labels for the vision-impaired. Experts can reliably evaluate this accessibility by applying a well-defined measurement procedure and web-oriented reporting like Access-I [L1]. However, the limited number of trained experts greatly restricts the ability to carry out such assessments at a large scale and keep the information up-to-date. On the other hand, dedicated social media apps like jaccede.com or wheelchair.org enable a crowdsourcing approach through the collection of partial information reported by different users, each with her own point-of-view, a method which may, however, lead to inconsistency. Both approaches are naturally complementary and can efficiently be combined through the use of open linked data techniques [2].

Although information and communication technologies are a clear enabler for physical accessibility, they also raise new obstacles in that the end-users then need to consult a computer or mobile interface to make sure that a place is physically accessible. The Web Accessibility Initiative (WAI) [L2] provides specific guidelines, such as the Web Content Accessibility Guidelines (WCAG) to help web developers make websites accessible while dynamic content is addressed through Accessible Rich Internet Applications (ARIA). Many useful evaluation and repair tools are also proposed, including for tailored and optimised usability and accessibility evaluation [4]. Many organisations propose practical labels that are based on WCAG, like AnySurfer in Belgium [L3].

Figure 1 graphically depicts the interplay between physical and digital accessibility, thus raising interesting questions about the current level of awareness about each kind of accessibility, the possible correlation between them, and what kind of synergies can further enhance user experience. In order to answer these questions, a preliminary study was carried out in Belgium by combining data from Access-I (for physical accessibility) and AnySurfer (for digital accessibility). Some methodological alignment was performed to define comparable notions of "unacceptable", "satisfactory" and "good" levels of accessibility for physical, sensory and mental impairments. The survey was conducted on places already assessed (so with some bias). Figure 1 shows a satisfactory level of physical accessibility in 60% of cases, of digital accessibility in 70% of cases and both in 33% of cases. The generally better level of digital accessibility, compared with physical accessibility, might be explained by the increased importance of digital presence and the lower effort required to achieve digital accessibility. Indeed, a website is easier and less expensive to repurpose than a building. The fact that only one third of the places studied currently achieve both dimensions means that these aspects are still largely being considered separately.

The current scope of both accessibility and digital assessments should also be reviewed: Physical accessibility is not limited to building boundaries but starts when the intention to travel to a place arises. And this scenario, eventually leading to the use of some service or equipment inside a building, relying on using a computer and/or a mobile interface like a smartphone or tablet to query the web about transportation, opening hours, and user specific accessibility. Physical and digital accessibilities then closely intertwine with retroaction loops illustrated in Figure 1.

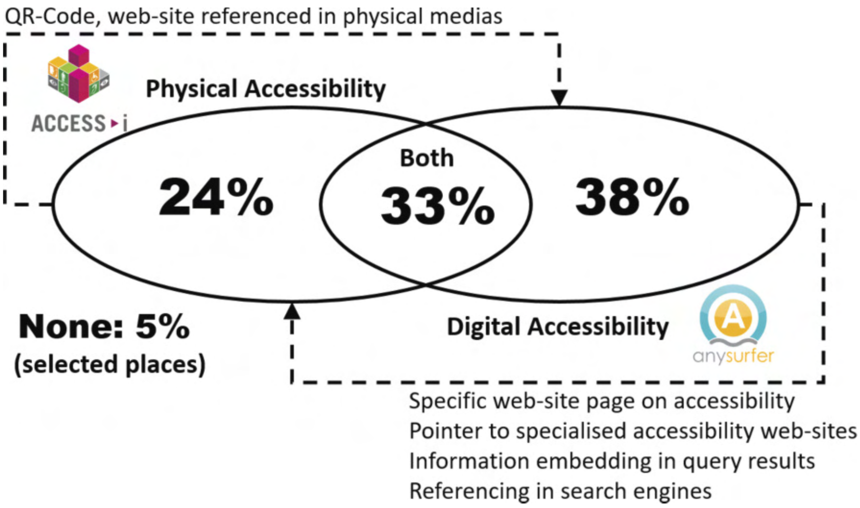

QR-Code, web-site referenced in physical medias
ACCESS▸i
Physical Accessibility
24%
Both
33%
38%
None: 5%
(selected places)
Digital Accessibility
anysurfer
Specific web-site page on accessibility
Pointer to specialised accessibility web-sites
Information embedding in query results
Referencing in search engines

Figure 1: The interplay between physical and digital accessibilities.

In addition to deepening our survey on the relations between physical and digital accessibility, we are also currently investigating how to effectively support accessibility using a digital mobile companion exploiting aggregated open accessibility data [3].

Links:

[L1] http://www.access-i.be
[L2] https://www.w3.org/WAI
[L3] http://www.anysurfer.be/en

References:

[1] C. Ponsard, V. Snoeck: "Unlocking Physical World Accessibility through ICT: A SWOT Analysis", Proc. of ICCHP'2014.

[2] C. Ponsard et al: "A Mobile Travel Companion Based on Open Accessibility Data", Proc. of ICCHP'2016.

[3] J. Vanderdonckt, A. Beirekdar: "Automated Web Evaluation by Guideline Review", Journal of Web Engineering, Vol. 4, No. 2, 2005, pp. 102-117.

Please contact:

Christophe Ponsard, CETIC, Belgium
+32 472 56 90 99, christophe.ponsard@cetic.be

Highly Sensitive Biomarker Analysis Using On-Chip Electrokinetics

by Xander F. van Kooten, Federico Paratore (Israel Institute of Technology and IBM Research – Zurich), Moran Bercovici (Israel Institute of Technology) and Govind V. Kaigala IBM Research – Zurich)

Highly sensitive and specific biochemical assays can provide accurate information about the physiological state of an individual. Leveraging microtechnology, we are developing miniaturised analytical tools for precise and fast biomolecular analysis. This work is being performed within the scope of the EU-funded project 'Virtual Vials' between IBM Research – Zurich and Technion – Israel Institute of Technology in Haifa.

In vitro diagnostics is a rapidly growing field and market that encompasses clinical laboratory tests and point-of-care devices, as well as basic research aimed at understanding the origin of diseases and developing improved testing methods and treatments. The global market for in vitro diagnostics is projected to grow to $75 billion by 2020. The development of novel tools that enable fast and sensitive analysis of biological samples with high throughput are essential for enabling discoveries and providing better care for patients.

In the past two decades, in vitro studies have shifted rapidly from classical tools to microfluidic devices (often also referred to as "lab-on-a-chip"). Microfluidic devices enable new capabilities and higher efficiencies in the analysis and control of biological liquid samples as many physical phenomena scale favourably as size is reduced. Effects such as diffusion and surface tension dominate the behaviour of fluids and solutes at micrometre length scales, while inertia and body forces such as gravity are negligible.

One of the key applications of in vitro diagnostics is the detection of biochemical species, such as proteins or nucleic acids that indicate a physiological condition or the presence of a disease. In many cases, the relevant analytes are present at such low concentrations that they cannot be detected using conventional methods. To overcome this, we make use of a specific electrokinetic separation and focusing technique called isotachophoresis (ITP). Isotachophoresis is a form of electrophoresis that is almost a century old, but has attracted renewed interest in the past decade, largely thanks to the strong growth of microfluidics research.

ITP uses a discontinuous buffer system to separate and concentrate target analytes based on differences in their electrophoretic mobility. As illustrated in Figure 1, the discontinuous buffer system consists of a leading (LE) and a terminating (TE) electrolyte, which respectively have a higher and a lower electrophoretic mobility than the analyte(s) of interest. When an electric field is applied, the analytes are focused at the continuously moving interface between the LE and TE. In this way, a "virtual vial" of several hundred picolitres is created, in which the concentration of analytes is enhanced by many orders of magnitude.

We focus on sensitive and rapid detection of proteins, which is key for early diagnosis of a large number of medical conditions. Protein analysis is commonly performed using surface immunoassays, which use antibodies immobilised on a surface to capture target proteins from a liquid sample. The number of captured proteins is converted into a readable signal by a variety of methods, such as electrochemical reactions, fluorescence or surface plasmon resonance. However, the performance of all these detection methods is limited by the kinetics of the protein-antibody reaction because the reaction time is inversely proportional to the concentration of proteins.

We use ITP to accelerate the surface reaction more than 1000-fold, so that a reaction that would require more than 20 hours under simple mixing can now be completed in under one minute. Figure 2a shows fluorescence images of the ITP-based immunoassay: At (t1) a fluorescent protein accumulates at the moving ITP interface; at (t2) the focused proteins reach a 100 µm long reaction site, where the accelerated reaction takes place, and at (t3) the proteins are carried away by ITP, leaving the reacted surface ready for detection. Combined with standard detection molecules available on the market, the technique can increase the sensitivity of existing immunoassays 1,300 times.

While ITP has proved to be a useful technique for accelerating reaction kinetics, the gains of this approach are ultimately limited by the volume of sample from which analytes are collected. In conventional microfluidic channels, ITP concentrates analytes from a sample volume of just a few

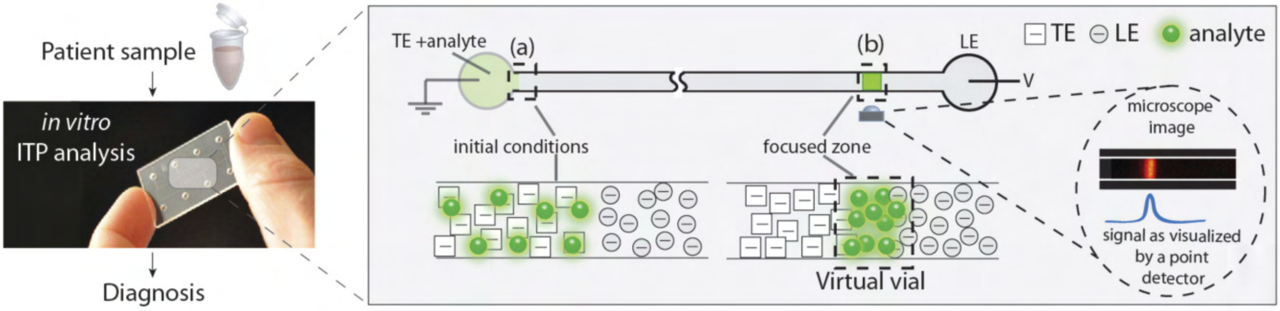


Figure 1: Schematic of ITP as a tool for in vitro diagnostics. Biochemical analytes from patient samples can be focused using ITP to increase their concentration locally. ITP uses a discontinuous buffer system to focus analytes at the interface between two electrolytes under an applied voltage V. This facilitates their detection and can lead to a faster and more sensitive diagnosis.

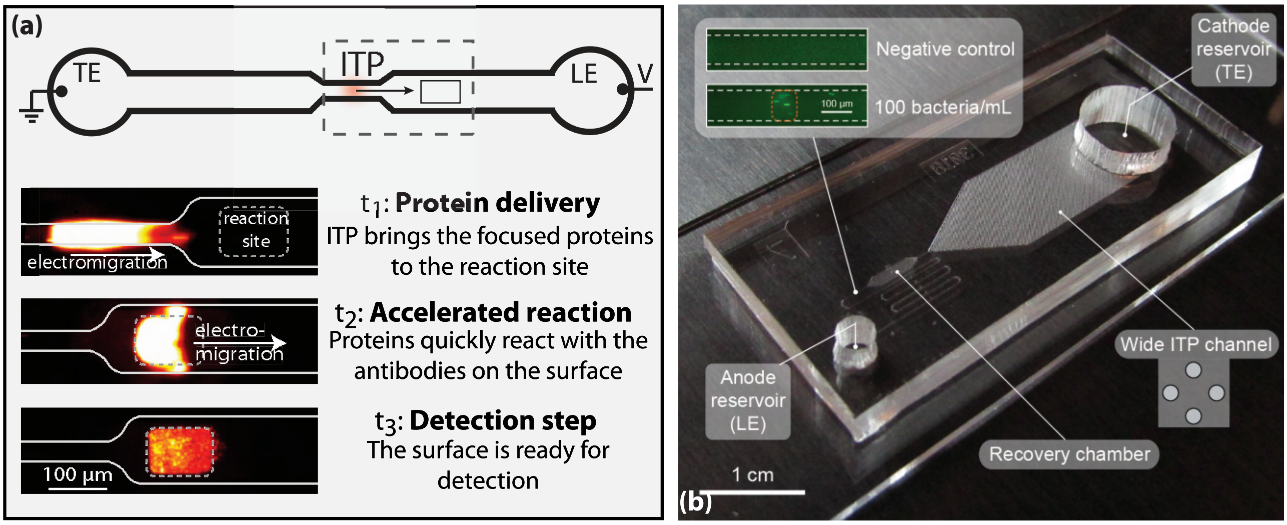


Figure 2: Two applications of ITP for in-vitro diagnostics we pursue. (a) In an ITP-based immunoassay, proteins are focused and delivered to a reaction site, where the enhanced concentration drives an accelerated reaction. (b) A device for ITP focusing from large sample volumes. The inset shows bacteria focused from an initial concentration of 100 bacteria/mL.

hundred nanolitres. At low concentrations, the presence of target molecules in the volume sampled by ITP becomes probabilistic, and multiple parallel tests are required to avoid false negatives.

As a step towards large-volume sample processing, we developed a chip that enables ITP focusing of biological analytes from 50 µL sample volumes in less than 10 minutes. As all analytes from this volume are focused into an ITP interface comprising just half a nanolitre, the final concentration factor is more than 100,000, which is one hundred times more than ITP focusing in conventional microfluidic chips.

The key to processing larger sample volumes lies in combining a wide-channel region (with a large internal volume) and a tapering channel. However, this gradual geometry change and the concomitant Joule heating resulting from local differences in the current density lead to undesired dispersion of the sample plug, which is detrimental to the concentration of the focused analyte and ultimately compromises detection downstream. Fortunately, this undesired dispersion can be mitigated by using a ‘“recovery chamber” at the end of the chip. This makes the design highly scalable, and the capability to process even larger sample volumes on passively cooled chips is limited only by the Joule heating in the channels. Therefore, if the chips are actively cooled, sample volumes of hundreds of microlitres may be processed in this way.

We believe these results are of fundamental significance for the in vitro diagnostics and research community as they suggest the use of assays for improving the sensitivity and speed of molecular analysis, and for processing large volume of samples on devices with microfluidic elements.

References:

[1] F. Paratore, et al.:”Isotachophoresis-based surface immunoassay”, Analytical Chemistry, 2017, 89 (14), pp. 7373–7381.

[2] X.F. van Kooten, M. Truman-Rosentsvit, G. V. Kaigala, and M. Bercovici: “Focusing analytes from 50 µL into 500 pL: On-chip focusing from large sample volumes using isotachophoresis, Scientific Reports 7: 10467, 2017.

Please contact:
Moran Bercovici, Israel Institute of Technology
mberco@technion.ac.il

Govind V. Kaigala, IBM Research – Zurich, Switzerland
gov@zurich.ibm.com

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

Dagstuhl Seminars and Perspectives Workshops

Schloss Dagstuhl – Leibniz-Zentrum für Informatik (LZI) is accepting proposals for scientific seminars/ workshops in all areas of computer science, in particular also in connection with other fields.

If accepted the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Due to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular details about event form and setup as well as the proposal form and the proposing process can be found on

http://www.dagstuhl.de/dsproposal.

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

Important Dates

- Proposal submission: October 15 to November 1, 2017
- Notification: End of January 2018
- Seminar dates: Between mid 2018 and mid 2019.

Event Report

A Conference for Advanced Students: IFIP TC6's AIMS

by Harry Rudin

IFIP's AIMS 2017, the 11th IFIP International Conference on Autonomous Infrastructure, Management and Security took place at the University of Zurich from July 10-13, 2017.

The idea driving the conference is special: AIMS is a single-track event targeted at junior researchers and PhD students in network and service management and security. It features a range of sessions including conference paper presentations, hands-on lab courses, and keynotes. The objective of AIMS is to offer junior researchers and PhD students a dedicated place where they can discuss their research work and experience, receive constructive feedback from senior scientists, and benefit from a number of practical hands-on sessions and labs on emerging technologies. By putting the focus on junior researchers and PhD students, AIMS acts as a complementary event in the set of international conferences in the network and service management community, providing an optimal environment for in-depth discussions and networking.

As with most IFIP TC6 conferences, the proceedings are available, free of cost, in the IFIP TC6 Open Digital Library: http://dl.ifip.org/db/conf/aims/aims2017/index.html

The proceedings, as a hard copy version, are also available in the Springer series Lecture Notes in Computer Science, Vol. 10356.

AIMS 2018 will take place in Munich in October 2018.

Link:
http://www.aims-conference.org/2017/

Call for Participation

W3C Workshop on WebVR Authoring: Opportunities and Challenges

Brussels, 5-7 December 2017

The primary goal of the workshop is to bring together WebVR stakeholders to identify unexploited opportunities as well as technical gaps in WebVR authoring. The workshop is targeted at people with experience of authoring WebVR content and people involved in the evolution of WebVR as a technology.

Topics

Participants will share good practices and novel techniques in creating WebVR-based content, and participate in breakout sessions and working discussions covering topics such as:

- Landscape of WebVR authoring tools;
- Creating and packaging 3D assets for WebVR;
- Managing assets for practical progressive enhancement;
- Progressive enhancement applied to the variety of user input in WebVR;
- Understanding and documenting WebVR constraints for 3D artists;
- Optimizing delivery of 360° videos to VR headsets on the Web;
- Practical approaches to building accessible WebVR experiences;
- Mapping the impact of ongoing evolutions of the Web Platform (Web Assembly, WebGPU, streams) on WebVR authoring;
- Impact of performance factors on authoring WebVR content;
- Creating convergence on WebVR advocacy platforms.

People interested in participation can express interest in attending the workshop or suggest a presentation through the web site. Attendance is free for all invited participants and open to the public, whether or not W3C members.

More information:
https://kwz.me/hLg

SICS becomes RISE SICS

The three Swedish institutes Innventia, SP and Swedish ICT, of which SICS was a part, have merged in order to become a stronger research and innovation partner for businesses and society: RISE Research Institutes of Sweden.

The vision of RISE is to be an internationally leading partner for innovation, offering applied research and development as well as industrialization and verification.

The business and innovation areas of RISE are:
- Mobility
- Energy and Bio-based Economy
- Life Science
- Digitalization
- Sustainable Cities and Communities

The ERCIM member institute SICS is changing its name to RISE SICS.

W3C Brings Web Developer Skills to the European Workforce

Since its creation in March 2015, W3C's e-learning platform W3Cx has trained more than half a million people from all over the world in Web design and development. By following W3Cx online courses, they have increased their digital skills that can lead to an exciting career in an in-demand and fast-growing field.

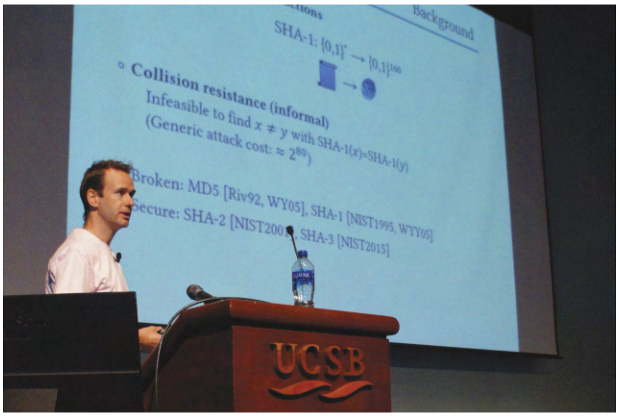
With this demand for Web development training, W3C has recently launched a "Front-End Web Developer" (FEWD) Professional Certificate on edX.org which consists of a suite of five W3Cx courses on the three foundational languages that power the Web: HTML5, CSS and JavaScript. Some of these courses have been developed in partnership with Microsoft, Intel and University Côte d'Azur.

The success of the W3Cx training programs underscores W3C's commitment in creating high quality courses that help learners to develop the critical skills and actionable knowledge needed for today's top jobs, and to develop the proficiency and expertise that employers are looking for.

W3C FEWD: https://www.edx.org/professional-certificate/front-end-web-developer-9

Awards for Breaking SHA-1 Security Standard in Practice

Researchers at CWI and Google won the Pwnie Award for Best Cryptographic Attack 2017 for being the first to break the SHA-1 internet security standard in practice in February. They received the prize on 26 July, during the BlackHat USA security conference in Las Vegas. The team consisted of Marc Stevens (CWI), Elie Bursztein (Google), Pierre Karpman (CWI), Ange Albertini and Yarik Markov (Google). The prize is a recognition for the most impactful cryptographic attack against real-world systems, protocols or algorithms, and the winners are selected by security industry professionals. The nominated read: "The SHAttered attack team generated the first known collision for full SHA-1. (…) A practical collision like this, moves folks still relying on a deprecated protocol to action."


Background
Collision resistance (informal)
Infeasible to find x ≠ y with SHA-1(x)=SHA-1(y)
Broken: MD5 [Riv92, WY05], SHA-1 [NIST1995, WY05]
Secure: SHA-2 [NIST200…], SHA-3 [NIST2015]
UCSB

Marc Stevens at the CRYPTO 2017 Best Paper Award lecture.

The CRYPTO 2017 Best Paper Award winning team. Photos source: Marc Stevens.

The research team also won the CRYPTO 2017 Best Paper Award on 22 August, during the CRYPTO 2017 conference in Santa Barbara. Although SHA-1 is deprecated, it is still used for digital signatures and file integrity verification, securing credit card transactions, electronic documents, GIT open-source software repositories and software distribution. On 17 August, Stevens and Daniel Shumow (Microsoft Research) presented an improved real-time SHA-1 collision detection at the USENIX Security conference in Vancouver, which is now used by default in Git, GitHub, Gmail, Google Drive, and Microsoft OneDrive.

More information: https://shattered.io.

ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Host of the World Wide Web Consortium.

Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
http://www.iit.cnr.it/

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
http://www.cwi.nl/

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
http://www.fnr.lu/

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
http://www.ics.forth.gr/

Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
http://www.iuk.fraunhofer.de/

INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
http://www.inria.fr/

I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
http://www.isi.gr/

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

SBA Research gGmbH
Favoritenstraße 16, 1040 Wien
http://www.sba-research.org/

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
http://www.sztaki.hu/

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
http://www.cs.ucy.ac.cy/

Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
http://www.mimuw.edu.pl/

VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
http://www.vttresearch.com