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

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

Copyright notice
All authors, as identified in each article, retain copyright of their work. ERCIM News is licensed under a Creative Commons Attribution 4.0 International License (CC-BY).

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/

Editorial Board:
Central editor:
Peter Kunz, ERCIM office (peter.kunz@ercim.eu)
Local Editors:
Austria: Erwin Schoitsch (erwin.schoitsch@ait.ac.at)
Belgium:Benoît Michel (benoit.michel@uclouvain.be)
Cyprus: Ioannis Krikidis (krikidis.ioannis@ucy.ac.cy)
France: Steve Kremer (steve.kremer@inria.fr)
Germany: Michael Krapp (michael.krapp@scai.fraunhofer.de)
Greece: Lida Harami (lida@ics.forth.gr),
Artemios Voyiatzis (bogart@isi.gr)
Hungary: Andras Benczur (benczur@info.ilab.sztaki.hu)
Italy: Maurice ter Beek (maurice.terbeek@isti.cnr.it)
Luxembourg: Thomas Tamisier (thomas.tamisier@list.lu)
Norway: Poul Heegaard (poul.heegaard@item.ntnu.no)
Poland: Hung Son Nguyen (son@mimuw.edu.pl)
Portugal: José Borbinha, Technical University of Lisbon (jlb@ist.utl.pt)
Spain: Silvia Abrahão (sabrahao@dsic.upv.es)
Sweden: Kersti Hedman-Hammarström (kersti@sics.se)
Switzerland: Harry Rudin (hrudin@smile.ch)
The Netherlands: Annette Kik (Annette.Kik@cwi.nl)
W3C: Marie-Claire Forgue (mcf@w3.org)

JOINT ERCIM ACTIONS

4 ERCIM Membership

5 ERCIM "Alain Bensoussan" Fellowship Programme

5 HORIZON 2020 Project Management

5 HRADIO – A New Project to Tackle Future of Internet Radio

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

8 In Codice Ratio: Scalable Transcription of Vatican Registers
by Donatella Firmani, Paolo Merialdo (Roma Tre University), Marco Maiorino (Vatican Secret Archives)

9 Teaching Archaeology to Machines: Extracting Semantic Knowledge from Free Text Excavation Reports
by Achille Felicetti (Università degli Studi di Firenze)

11 MTAS – Extending Solr into a Scalable Search Solution and Analysis Tool on Multi-Tier Annotated Text
by Matthijs Brouwer (Meertens Institute)

12 Phonetic Search in Audio and Video Recordings
by Ioannis Dologlou and Stelios Bakamidis (RC ATHENA)

13 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)

14 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)

16 ContentCheck: Content Management Techniques and Tools for Fact-checking
by Ioana Manolescu (Inria Saclay and Ecole Polytechnique, France)

17 Argumentation Analysis of Engineering Research
by Mihály Héder (MTA-SZTAKI)

Information Management Strategy

18 A Back Bone Thesaurus for Digital Humanities
by Maria Daskalaki and Lida Charami (ICS-FORTH)

19 Meeting the Challenges to Reap the Benefits of Semantic Data in Digital Humanities
by George Bruseker, Martin Doerr and Chryssoula Bekiari (ICS-FORTH)

21 HumaReC: Continuous Data Publishing in the Humanities
by Claire Clivaz, Sara Schulthess and Anastasia Chasapi (SIB)

22 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)

23 Trusting Computation in Digital Humanities Research
by Jacco van Ossenbruggen (CWI)

Research Infrastructure Development

25 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)

26 DIGILAB: A New Infrastructure for Heritage Science
by Luca Pezzati and Achille Felicetti (INO-CNR)

28 Building a Federation of Digital Humanities Infrastructures
by Alessia Bardi and Luca Frosini (ISTI-CNR)

29 Knowledge Complexity and the Digital Humanities: Introducing the KPLEX Project.
by Jennifer Edmond and Georgina Nugent Folan (Trinity College Dublin)

3D Analysis Techniques

30 **Restoration of Ancient Documents Using Sparse Image Representation**
by Muhammad Hanif and Anna Tonazzini (ISTI-CNR)

32 **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)

33 **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)

35 **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)

36 **Physical Digital Access Inside Archaeological Material**
by Théophane Nicolas (Inrap/UMR 8215 Trajectoires), Ronan Gaugne (Université de Rennes 1/IRISA-Inria), Valérie Gouranton (INSA de Rennes/ IRISA-Inria) and Jean-Baptiste Barreau (CNRS/CReAAH UMR 6566)

Information Visualization and Communication

37 **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 (LIST)

39 **Cultural Opposition in former European Socialist Countries: Building the COURAGE Registry**
by András Micsik , Tamás Felker and Balázs Nász (MTA SZTAKI)

40 **Locale, an Environment-Aware Storytelling Framework Relying on Augmented Reality**
by Thomas Tamisier, Irene Gironacci and Roderick McCall (Luxembourg Institute of Science and Technology)

42 **The Biennale 4D Project**
by Kathrin Koebel, Doris Agotai, Stefan Arisona (FHNW) and Matthias Oberli (SIK-ISEA)

43 **The Clavius Correspondence: From Digitization to Visual Exploration of Knowledge**
by Matteo Abrate, Angelica Lo Duca, Andrea Marchetti (IIT-CNR)

45 **Service-oriented Mobile Social Networking**
by Stefano Chessa and Michele Girolami (ISTI-CNR)

RESEARCH AND INNOVATION

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

46 **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)

48 **Distributional Correspondence Indexing for Cross-Lingual and Cross-Domain Sentiment Classification**
by Alejandro Moreo Fernández, Andrea Esuli and Fabrizio Sebastiani

49 **Measuring Bias in Online Information**
by Evaggelia Pitoura (University of Ioannina), Irini Fundulaki (FORTH) and Serge Abiteboul (Inria & ENS Cachan)

50 **The Approximate Average Common Submatrix for Computing the Image Similarity**
by Alessia Amelio (Univ. of Calabria)

52 **My-AHA: Stay Active, Detect Risks Early, Prevent Frailty!**
by Nadine Sturm, Doris M. Bleier and Gerhard Chroust (Johanniter Österreich)

53 **A New Vision of the Cruise Ship Cabin**
by Erina Ferro (ISTI-CNR)

54 **Trend Analysis of Underground Marketplaces**
by Klaus Kieseberg, Peter Kieseberg and Edgar Weippl: (SBA Research)

55 **RDF Visualiser: A Tool for Displaying and Browsing High Density of RDF Data**
by Nikos Minadakis, Kostas Petrakis, Korina Doerr and Martin Doerr (FORTH)

57 **Services for Large Scale Semantic Integration of Data**
by Michalis Mountantonakis and Yannis Tzitzikas (ICS-FORTH)

58 **On the Interplay between Physical and Digital World Accessibility**
by Christophe Ponsard, Jean Vanderdonckt and Lyse Saintjean (CETIC)

60 **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)

EVENTS, IN BRIEF

Call for Proposals

60 **Dagstuhl Seminars and Perspectives Workshops**

Report

60 **A Conference for Advanced Students: IFIP TC6's AIMS**

62 **W3C Workshop on WebVR Authoring: Opportunities and Challenges**

In Brief

63 **SICS becomes RISE SICS**

63 **W3C Brings Web Developer Skills to the European Workforce**

63 **Awards for Breaking SHA-1 Security Standard in Practice**

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

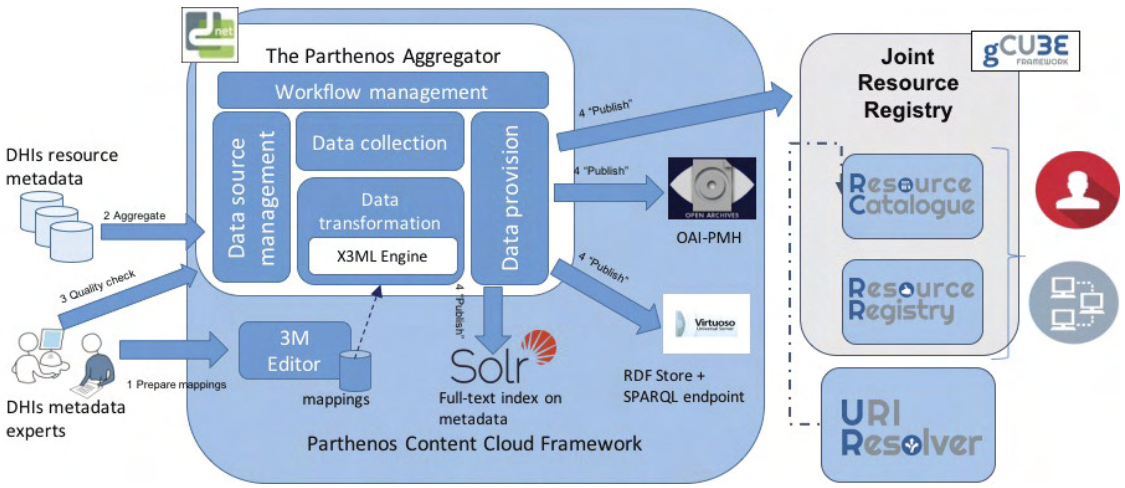

Figure 1: Technical framework for DHIs federation.

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];