Number 100 January 2015

ERCIM NEWS

www.ercim.eu

Special theme: What is Computation? Alan Turing's Legacy
Special theme: Smart Energy Systems
Special theme: Mobile Computing
Special theme: Intelligent Cars
Special theme: Image Understanding
Special theme: Linked Open Data
Special theme: Cyber-Physical Systems
Special theme: Smart Cities
Special theme: Software Quality
Special theme: Cybercrime and Privacy Issues
Special theme: Big Data
Special theme: Evolving Software
Special theme: Ambient Assisted Living
Special theme: ICT for Cultural Heritage
Special theme: Unconventional Computing Paradigms
Special theme: Intelligent and Cognitive Systems
Special theme: Cloud Computing Platforms, Software and Applications
Special theme: Computational Biology
Mathematics for Everyday Life
Supercomputing at Work
Safety-Critical Software
Towards Green ICT
Digital Preservation
Special theme: Simulation & Modelling
Special: The Future WEB
Special: Machine Perception
Special: Analysis, Diagnosis, Planning and Simulation of Industrial Systems
Special: Games Technology
Special: Automated Software Engineering
Special: Grids: The Next Generation
Special: Biomedical Informatics
Special: Environmental Modelling
Special: Multimedia Informatics
Special: Security and Trust Management
Special: Cognitive Systems
Special: Embedded Systems
Special: Semantic Web
ERCIMathematics
Information Security
e-government
Ambient Intelligence
Control and System Theory
Special Theme: Web Technologies
Special Theme: Bioinformatics
Special Theme: Computer Graphics and Visualization

Special theme:

Scientific Data sharing and re-use

Also in this issue:

Keynote:
by Carlos Morais, European Commission

Joint ERCIM Actions:
100 issues of ERCIM News
ERCIM 25 Years Celebration

Research and Innovation:
Simulations Show How Lightning Creates Antimatter

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: Jérôme Chailloux
ISSN 0926-4981

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

Copyright notice

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

ERCIM News online edition
The online edition is published at http://ercim-news.ercim.eu/

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/

Next issue
April 2015, Special theme: The Internet of Things & the Web of Things

Contents

JOINT ERCIM ACTIONS

ERCIM White Papers

KEYNOTE

SPECIAL THEME

The special theme section “Scientific Data Sharing and Reuse” has been coordinated by Costantino Thanos, ISTI-CNR and Andreas Rauber, TU Vienna.

Introduction to the Special Theme

RESEARCH AND INNOVATION

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

EVENTS, IN BRIEF

Reports

Announcements

In Brief

100 issues of ERCIM News

Our First 25 Years: 100 Issues of ERCIM News

by Peter Kunz, ERCIM News editor

With this edition, ERCIM is celebrating the 100th issue of its magazine, ERCIM News. Since ERCIM's creation in 1989, the quarterly ERCIM News has been reporting on leading edge European research and developments in Information and Communication Science and Technology (ICST) and Applied Mathematics.

When Inria, CWI and former GMD founded ERCIM in 1989, the establishment of an 'in-house magazine' with the aim of reporting on joint activities was one of the first 'joint actions'. With new members joining ERCIM, our founders Henk Nieland (CWI), Siggi Münch (GMD) and Laure Reinhart (Inria), decided to establish an editorial board with one representative per member institute, responsible for collecting contributions from their institute. ERCIM rapidly evolved from a black-and-white in-house magazine to a full colour publication covering reports and news about scientific projects from all over Europe and even beyond.

In the early nineties, the Internet entered the research institutes and soon raised the question about an on-line version of ERCIM News. By October 1994, the newsletter was published both in printed and electronic format. At that time, ERCIM News was among the first 5,000 Web sites in the world. Surprisingly, the electronic edition did not detract from the success of the printed edition – instead, many new readers who found us on the Web also subscribe to the printed edition, thus increasing its circulation. The peak was reached in 2009 with a circulation of over 10,000 printed copies. Since then, presumably with the spread of smart phones and tablets, the circulation of the printed edition has reduced whilst the online readership has increased. In 2012, for the first time, more people subscribed to the electronic than to the printed edition. ERCIM News currently maintains a circulation of 4,500 printed copies and more than 7,000 people subscribed to the online edition.

From the early issues on, each issue has focused on a special theme identified by the editorial board. The ERCIM News series has thus become a unique collection providing an overview on a wide range of research topics in ICST and Applied Mathematics. All articles in ERCIM News are written by the scientists themselves and professionally edited. The structure of the articles and the limited length also make them comprehensible for non-experts. Thanks to these unique characteristics, ERCIM News has become well-known in the world of scientific publications, and regular positive feedback from our readers has encouraged us to continue in this way. Indeed, our readership comprises not only scientists, but also students, decision makers, professionals within the industry, representatives from the European Commission, and politicians.

The last quarter of a century has seen tremendous advances in technology, especially with the Internet revolution. ERCIM News has witnessed this evolution by focusing on current and emerging topics for the special themes. We have seen completely new areas emerging, for instance, the birth and development of the World Wide Web, to which we dedicated special themes in April 1996 ("The World Wide Web"), in April 2000 ("Web Technologies"), October 2002 ("Semantic Web"), January 2008 ("The Future Web), January 2009 ("The Sensor Web"), April 2009 ("Future Internet Technology") and the forthcoming issue in April 2015 ("The Web of Things"). Other areas have been emerging as increasing computing power enables new fields of applications: in 1993, for example, we had the special theme on "Parallel Architectures", showing the progress made in the use of high-performance computing systems. Many research areas on which we reported later, would not have been possible without the research carried out at that time. Other research topics remain a challenge over the years. Software Quality, for example, was the theme of an issue in July 1992, and remains a very important issue today (see ERCIM News no. 99, October 2014).

For each edition, ERCIM News also invites a personality to write a keynote statement relevant to the European scientific community. Keynote authors have included: Tim Berners-Lee, the inventor of the Web, Franco Malerba and Elly Plooij-van Gorsel, both Members of the European Parliament, the European Commissioners Erkki Liikanen, Philippe Busquin and Viviane Reding, UK Minister Lord Sainsbury, the Irish Prime Minister Bertie Ahern, and Eberhard van der Laan, Mayor of Amsterdam.

All previous issues are available on the ERCIM News website. Today, ERCIM News is published in print, and electronically in PDF, epub format and in HTML.

One hundred issues of ERCIM News means more than 2,000 published articles. The popularity of ERCIM News can be credited primarily to our authors to whom the ERCIM editorial board wants to express their warmest thanks on this occasion.
http://ercim-news.ercim.eu

Members of the ERCIM Editorial Board.

Messages from our readers

Thank you to the many readers who sent us a message on the occasion of the 100th issue of ERCIM News. Here is a small selection of the messages received.

"Congratulation to your 100th issue of ERCIM News! I favor ERCIM News because of its novelty and excellent style of reporting on actual science and development."

R. Dillmann, Karlsruhe Institute of Technology (KIT)

"Well done on reaching this amazing level of output - I look at any ERCIM News that links with my current work and often ones that don't just out of interest!"

J. Crowcroft, Cambridge University

"ERCIM News contains stimulating information. Thank you for presenting informed articles describing the big EU initiatives. I look forward to each ERCIM issue and I often share them with colleagues.
Congratulations on your anniversary!"

J. Peasant, U.S. Air Force

"Great work, some videos would be nice though."

A. Cartagena Gordillo, UCSP - UCSM, Perú

"100 issues, more than 1000 research contributions, sure more than 100000 of new ideas for research.
Congratulations and thanks for scientific advance."

E. Arias, University of Castilla-La Mancha

"Dear ERCIM Editors, please accept my congratulations on your excellent and so useful work. On the occasion of the 100th issue, let me wish you the next hundred."

*L. Polkowski, Polish Japanese Institute IT,
Warszawa, Poland*

"Congratulations for the previous 99 issues of ERCIM News. It would be interesting in issue 101 to make some forecasts about what will be important in issue 200."

J. Falcão e Cunha, INESC

"Best wishes! ERCIM News is a really good European pendant to the American COMMUNICATIONS of the ACM. Please go ahead!"

M. Pieper, Fraunhofer FIT

Domenico Laforenza, President of ERCIM

Alberto Sangiovanni Vincentelli

Carlo Ratti

ERCIM 25th Anniversary Celebration

On the occasion of ERCIM's 25th anniversary, a special session and panel discussion took place on 23 October in the auditorium of the CNR Campus. Speakers and representatives from research, industry, the European Commission, and the ERCIM community presented their views on research and future developments in information and communication science and technology (ICST).

Domenico Laforenza, President of ERCIM welcomed more than 100 ERCIM delegates, ERCIM postdoc fellows and guests to the special anniversary session and symposium. He gave a brief overview on ERCIM's history, achievements and outlook.

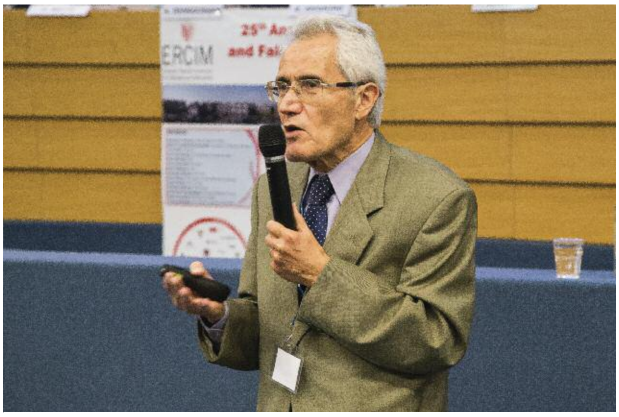
Alberto Sangiovanni Vincentelli, Professor at the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley gave insight in cyber-physical systems with his talk "Let's get physical: marrying computing with the physical world".

Carlo Ratti, Director of the Senseable City Lab at the Massachusetts Institute of Technology (MIT), USA described some of the Lab's projects such as exploring the spatio-temporal patterns of mobile phone activity in cities, and "Ambient Mobility", a smart mobility research project to analyze, design, develop and test senseable and sustainable mobility systems in interaction with other future urban systems. "Ambient Mobility" is a joint initiative between the Senseable City Lab at MIT and the Mobility Innovation Lab at the Fraunhofer Institute for Industrial Engineering IAO.

Alain Bensoussan, former President of Inria and one of the co-founders of ERCIM, now Professor of Risk and Decision Analysis at the School of Management at the University of Texas at Dallas, discussed the question: "Big data and big expectations: Is a successful matching possible?"

The following two speakers introduced the latest ERCIM White Papers. Rigo Wenning, W3C presented the White Paper "Security and Privacy Research Trends and Challenges". The White Paper "Big Data Analytics: Towards

The audience



Alain Bensoussan

The panelists. From left: Carlo Ghezzi, Thomas Skordas, Paola Inverardi, Jean-Pierre Bourguignon, Domenico Laforenza and Fabio Pianesi.

a European Research Agenda", was presented by Fosca Giannotti, ISTI-CNR (see also articles on pages 9 and 10).

Radu Mateescu, senior researcher at Inria Grenoble - Rhône-Alpes and chair of the ERCIM Working Group on Formal Methods for Industrial Critical Systems then presented the achievements of two decades of this ERCIM Working Group which has achieved world-wide recognition in this field.

The presentations were followed by a panel discussion on the topic "ICT Research in Europe: How to reinforce the cooperation between the main actors and stakeholders". The panel was moderated by Carlo Ghezzi, President, Informatics Europe (Politecnico di Milano, Italy). The participants were Domenico Laforenza, President, ERCIM AISBL (CNR), Fabio Pianesi, Research Director, ICT Labs, European Institute for Innovations and Technology, Jean-Pierre Bourguignon, President of the European Research Council (ERC), Paola Inverardi, ICT Italian Delegate, Rector of the University of L'Aquila, Italy and Thomas Skordas, head of the FET Flagships Unit, European Commission, DG Communications Networks, Content and Technology (DG CNECT).

Juan Reutter, Assistant Professor at the Department of Computer science at the Catholic University of Chile, received the 2014 ERCIM Cor Baayen Award from Cor Baayen, ERCIM's president d'honneur himself and closed the session with a talk about his work, which deals mainly with foundations of data management (see article on page 8).

In conjunction with the anniversary event, CNR also hosted the fourth ERCIM ABCDE Fellowship Seminar and the regular ERCIM Fall meetings. The Fellowship Seminar was a two-day event for postdoctoral ERCIM fellows hosted in the frame of the ERCIM ABCDE fellowship programme, supported by the FP7 Marie Curie Actions - People, Co-funding of Regional, National and International Programmes (COFUND). 20 fellows participated in the seminar.

Cor Baayen

http://www.ercim.eu/25years

Participants at the gala dinner.

From left: Domenico Laforenza, Alain Bensoussain, Cor Baayen, Keith Jeffery and Michel Cosnard.

Juan Reutter Receives the 2014 Cor Baayen Award

Juan Reutter, Assistant Professor at the Department of Computer Science at the Catholic University of Chile, has received the 2014 ERCIM Cor Baayen Award. The Cor Baayen Award is given annually by ERCIM to a promising young researcher in the field of Informatics and Applied Mathematics.

Juan's main field of interest is foundations of data management. His PhD thesis entitled "Graph Patterns: Structure, Query Answering and Applications in Schema Mappings and Formal Language Theory'', prepared at the University of Edinburgh under the supervision of Leonid Libkin, is in the area of graph data management. The main goal of Juan's thesis was the study of graph patterns and querying mechanisms for them. His work provided a full classification of the expressive power of graph patterns based on regular expressions. One of the key conclusions of the thesis is that querying graph patterns is computationally harder than querying patterns in other paradigms, such as relational or XML databases, which has implications, for example, for applications that need to integrate or transform graph data. The computational cost of querying patterns also necessitates the search of tractable restrictions and efficient heuristics. His dissertation shows a connection between querying graph patterns and constraint satisfaction, pointing to opportunities to employ many heuristics developed in that field.

Juan Reutter receives the Cor Baayen Award

During his PhD, Juan received several awards. For his work in exchanging knowledge bases, he received the Best Paper Award at PODS 2011 (a top database conference) and the "Ramon Salas" award for the best work in engineering produced by Chilean researchers in 2012. In his short career Juan has already produced an impressive number of publications and research results in different topics. The list of his publications includes nine journal papers, eleven conference papers, two book chapters and four workshop papers. His work was published in several top-rated conferences and journals, including two journals of the ACM.

Following the completion of his PhD, Juan continued his work in graph data management. He is now studying the problem of static analysis and optimization of graph queries. To pursue this line of research, Juan received a competitive start-up fund for a project called "Query Languages and Views for Graphs with Data", awarded in 2013 by the National Commission for Scientific and Technological Research in Chile. He is also an associate investigator of the "Millennium Nucleus Centre for Semantic Web Research", a competitive research project involving three Chilean universities that is funded by the Chilean Ministry of Economy. Juan has also been working in the field of schema mappings (exchanging data between applications using different schemas, according to a set of predefined translation rules). His most significant contribution is a theoretical framework for studying the exchange of complex databases such as knowledge bases or databases containing ontologies. This work has also had implications outside database theory, and has recently been cited by several papers in the area of description logics.

2014 Cor Baayen Award:

Winner:
- Juan Reutter, Pontifica Universidad Católica de Chile

Honorary mention:
- Marc Stevens, CWI

Other finalists:
- Ahmad Al-Shishtawy, SICS Swedish ICT
- Balint Antal, University of Cambridge and University of Debrecen
- Alberto Bacchelli, Delft University of Technology
- Antonio Filieri, University of Stuttgart
- Laszlo Gyongyosi, Budapest University of Technology and Economics, Hungarian Academy of Sciences
- Michal Havlena, ETH Zurich
- Antti Hyttinen, California Institute of Technology
- George Kousiouris, ICCS/NTUA
- Anne-Cécile Orgerie, CNRS
- Samir Perlaza, Inria
- Shahid Raza, SICS Swedish ICT
- John Vardakas, Iquadrat Informatica S.L.

The Cor Baayen Award is awarded each year to a promising young researcher in computer science and applied mathematics. The award was created in 1995 to honour the first ERCIM President.

http://www.ercim.eu/activity/cor-baayen-award

ERCIM White Papers

ERCIM launched an initiative to identify emerging grand challenges and strategic research topics in Information and Communication Science and Technology (ICST).

In 2014, two initial Task Groups were formed to investigate topics related to “Big Data Analaytics” and “Cyber-Security and Privacy”respectively. Further topics are currently being identified. The groups were composed of researchers representing a broad cross-section of interests, affiliations, seniority, and geography. Their results are published in two White Papers, which are briefly presented in the following two articles.

The full version of both White Papers is available for download from the ERCIM web site at
http://www.ercim.eu/publications/strategic-reports

ERCIM White Paper

Big Data Analytics: Towards a European Research Agenda

Editors: Mirco Nanni, Costantino Thanos, Fosca Giannotti and Andreas Rauber

The information hidden within big data may provide the key to solving many problems within society, business and science. However, turning an ocean of messy data into knowledge and wisdom is an extremely challenging task.

In this paper, we outline our vision of big data analytics in Europe, based on the fair use of big data with the development of associated policies and standards, as well as on empowering citizens, whose digital traces are recorded in the data. The first step towards this objective is the creation of a European ecosystem for big data analytics-as-a-service, based on a Federated Trusted Open Analytical Platform for Knowledge Acceleration. The goal is to yield a data and knowledge infrastructure that offers individuals, scientists, institutions and businesses: (i) access to data and knowledge services, and (ii) access to analytical services and results within a framework of policies for access and sharing based on the values of privacy, trust, individual empowerment and public good. Several requirements need to be fulfilled at four distinct levels as discussed in detail below:

- *Scientific and technological challenges.* Solutions to difficult problems are needed, including: i) the development of new foundations for big data analytics, which integrate knowledge discovery from big data with statistical modelling and complex systems science, ii) semantics data integration and enrichment technologies, which make sense of big data and make them usable for high-level services, iii) scalable, distributed, streaming big data analytics technologies to master the intimidating volume and speed of big data.

- *Data requirements.* The potential value that lies within big data can only be unleashed if a proper, efficient, fair and ethical access to data is provided to the relevant actors. This poses many techno-social questions: who owns and who may use personal data? What is the real value of such data? How can different data sources be accessed and linked? How can individuals be empowered through the capability to access, use, handle and control the usage of, their own data? How can we communicate to individuals and institutions the social and/or economic impact of personal data? How can open data initiatives be boosted and federations of linked data developed?

- *Education and data literacy*. Acquiring wisdom and value from big data requires competent data analytics professionals with considerable expertise in statistics and machine learning. Skills must be developed on how to exploit data and their analysis to develop successful business initiatives. Moreover, given the pervasiveness of big data in most disciplines of human knowledge and research, elements of data science should be provided to students at all levels of education, from high-schools to university curricula.

- *Promotional initiatives for data analytics and Big Data Analytics-as-a-service-as-a-service.* In order for Big Data Analytics-as-a-service to flourish in Europe, we must have a strategy to promote development along some key directions. These include: supporting the creation of big data analytics centres that are accessible to researchers, public administrations, and medium and small companies; incentivizing the adoption of a layered framework to increase interoperability across the single data repositories; and European leadership in the development of privacy-enabling solutions.

Finally, a very effective way of promoting and helping the development of big data analytics is to create successful, large-scale showcases in high-impact application domains. Some particularly effective examples might include: smart cities and communities; big data analytics for developing countries; the management of the global market of jobs; the (quantitative) assessment of results of European projects and activities; and the development of Big Data-aware Official Statistics.

Final recommendations

Over the last 10 years, European research has invested significant capital in database and data mining technology, and has developed a strong base of expertise and innovation in these areas. Future actions should capitalize and advance on this base. To this end, our recommendations are:

- The EU should spawn federations of key public and private actors, in challenging multidisciplinary domains to provide a critical mass for starting up Federated Trusted Open Analytical Platforms for Knowledge Acceleration and creating incentives for further actors to join.

- The EU should support the creation of big data analytics centres accessible to researchers, public administrations, and medium and small companies.
- Funding and supporting the development of the technologies needed to empower citizens, public institutions and businesses based on the values of the Federated Trusted Open Analytical Platforms for Knowledge Acceleration.
- Promoting the development of a normative framework for the above mentioned empowering of citizens, public institutions and businesses along four dimensions: privacy-preservation, trust management, individual empowerment and public good.
- Promoting education of novel data scientists and 'datacy'.
- Promoting incentives for providing data access to researchers, businesses and public administrations. Examples include: assigning rewards to and/or facilitating virtuous business actors that share and maintain open data portals; giving value to 'data citation' in research, i.e. recognizing the citations to a data collection as a valid bibliometrics indicator; and enforcing regulations.

Organization and methodology

A group of twenty researchers from the areas of core database technology and data analytics met in Pisa, Italy in May 2014 to discuss the state of research in these areas in the context of big data, its impact on practice or education, and important new directions. The attendees represented a broad cross-section of interests, affiliations, seniority, and geography.

Before the workshop, each participant submitted a short paper, to be shared among participants, summarizing his/her vision of big data analytics over the next 10 years. During the workshop two subgroups were formed: one focused mainly on data management issues introduced by big data and the other on the new analytical opportunities opened by big data. The two groups alternated between separate meetings/brainstorming sessions and plenary meetings to share results and keep the overall efforts focused. The workshop terminated with the collaborative preparation of a declaration of aims and an index of context that were later expanded into the present paper.

Link:
The White Paper is available for download at
http://www.ercim.eu/images/stories/pub/white-paper-BigDataAnalytics.pdf

Please contact:
Mirco Nanni, ISTI-CNR,
Tel: +39 050 6212843
E-mail: mirco.nanni@isti.cnr.it

ERCIM White Paper

Cyber-Security and Privacy

Editors: Javier Lopez, Fabio Martinelli and Pierangela Samarati

Cyber-security and privacy are very active areas of research, and experts in this field are in growing demand by companies. Research in this area is highly relevant in today's world - many aspects of our daily life depend on ICT - from mobile phones (more than one billion devices currently have the same operating system), computers in offices and programmable machines in factories, and intelligent surveillance cameras contributing to our safety (and infringing on our privacy). The pervasiveness of ICT increases the potential attack surface available to attackers, expanding the opportunities and potential for damage. Both the number of attackers and motivations for attacks are on the increase, indicating that cyber-security attacks can be a prosperous business.

Cyber-security is a long-standing research topic with many success stories represented in the literature and in the industry standards and products. Yet, cyber-attacks are increasing and their impacts becoming increasingly significant. There is no doubt that this is related to the expansion of the role of ICT and the fact that threats evolve along with technology.

Cyber-attacks receive regular media coverage, thus raising awareness and concern about the issue within society. The protection of personal information is also a key concern. Many consider this to be a lost battle already, since most of our everyday life can be monitored by private companies (in addition to states). It should not be necessary, however, to forgo one's privacy in order to enjoy technology. Work is needed to enable users to benefit from technological advances without compromising on privacy. Technological solutions should be developed to empower users with full control over their own data as well as to provide technological support to assist in regulating the protection of data.

This document identifies research areas that could provide significant contributions to reduce cyber-insecurity and where ERCIM Institutions have significant research and innovation capabilities able to drive cooperative research efforts in the field. The following areas were identified: system security; security engineering; security operation management; data protection; security and big data; access control; quantitative aspects of security; practical and usable security; cryptography; trust management systems; digital freedom; network security.

In particular, we make the following recommendations for research activities in cyber security and privacy:

- Research activities are necessary in the area of system security in particular related to malware detection and protection as well as to security and privacy aspects of social networks.
- Another relevant area for security is related to the secure engineering processes and methodologies. Significant

emphasis should be placed on the *-by-design aspects (where * stands for security, privacy, trust, etc).
- In addition, attention should be devoted to operational security. In particular, it is important to improve the qualification of the personnel, applied methods for cyber attack detection and response, and sharing the gained information and experience in public/private partnerships.
- More research is needed to ensure proper protection of data controlled by a third party (social network, cloud provider, outsourcer, etc.) during transmission, processing or storage.
- Big data analysis also imposes novel challenges on security research. On the one hand, the analysis itself should be secure and privacy-aware, on the other, the analysis can be used to improve security of systems and services.
- A distributed world requires efficient adaptation of well-known technologies such as access control. Access control models should become more flexible, rely less on dynamic identities, and use advance solutions for efficient key management.
- Research activities are required to develop more precise measuring in cyber security. These measurements should be reliable and realistic, supporting decision-makers at different levels: from low level technical personnel setting specific security controls properly to high level managers considering risk mitigation strategies for the whole system.
- Research is also needed to improve the usability of security by understanding the behaviour of users and developing controls which are easy to apply. User-centric research should also reduce social engineering attacks.
- Cryptography is another high profile area, which is still of paramount importance. It is important that we pay attention not only to the most advanced directions, i.e., post-quantum cryptography, but also to the correct and trustable implementation of classical methods.
- Research activities are also needed to adapt and enhance trust management. The activities should consider establishing trust in a dynamic environment, where identities may be unknown, only few anchors of trust exist, and different trust models should interact.
- More research is required into enforcing the digital rights of individuals: protecting their privacy; empowering users in the management of their own data; balancing user identification and anonym usage and similar issues.
- Research is required into better protection of established networks, by protection of internal communications, secure interaction with other networks, and ensuring vulnerability-free design.

We also considered the following more general concerns :
- The technical and legal aspects of security are interdependent, and collaboration between researchers and practitioners of both areas should be established from the outset of any project.
- Reporting of cyber incidents should be made easier, since such events may impact the well-being of citizens. Such reports should be obligatory and provide enough information for a thorough analysis of causes and consequences.
- Last but not least, activities are required into raising public awareness of all aspects of cyber-security and privacy.

Methodology

In accordance with the initial appointment of the ERCIM BoD, the white paper was put together based on the outcomes of expert group workshops and consensus building. Thus, we planned the focus expert group workshop on 9 September 2014, prior to the ERCIM STM WG meeting (10-11 September 2014) in Wroclaw, in cooperation with the 19th European Symposium on Research in Computer Security (ESORICS). Before the workshop, members of the expert group supplied short position statements. Based on the position statements and the follow-up discussion, a set of relevant technological areas were identified, together with application domains and other non-technical issues. The initial findings of the expert group were presented during the ERCIM STM workshop, where we received additional feedback on the research challenges to be addressed. Follow-up meetings were carried out via teleconference between the expert group members. The white paper was then collaboratively prepared by the expert group members (about 30, with an even split between research and industry expertise).

Link:
The White Paper is available for download at
http://www.ercim.eu/images/stories/pub/white-paper-STM.pdf

Please contact:
Fabio Martinelli, IIT-CNR, Italy
Tel: +39 050 3153425
E-mail: Fabio.Martinelli@iit.cnr.it

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-month duration, spent in one ERCIM member institute. Fellowships are proposed according to the needs of the member institutes and the available funding.

Conditions

Applicants must:
- have obtained a PhD degree during the last eight years or be in the last year of the thesis work with an outstanding academic record
- be fluent in English
- be discharged or get deferment from military service
- have completed the PhD before starting the grant.

Application deadlines

30 April and 30 September

More information and application form:
http://fellowship.ercim.eu/

e-Infrastructures Enabling Trust in the Era of Data-Intensive Science

Carlos Morais Pires, Scientific Officer at the European Commission, Excellence in Science DG/CONNECT. Carlos Morais Pires, coordinates the area of "Scientific Data e-Infrastructures" at the European Commission, DG CONNECT.

The World Wide Web is a global communication platform transforming research and education. The data-driven paradigm, although in different contexts, affects all fields of science. New scientific methods are emerging supported by unprecedented ability to move data around and the capacity to process them even in extreme large volumes . Trust in the scientific enterprise builds on evidence-based methods and mechanisms of peer review and scrutiny. This has been working well for centuries involving more or less homogeneous groups of scientists. But if trust is a fundamental and time invariant value of science, it has to scale in order to preserve it in a hyper connected world. It has to take into account multidisciplinary approaches, citizens' growing scientific literacy and their engagement in science. The Web obliges us to reflect and put in place a framework for webs of trust. In order to scale, a trust-enabling framework has to get the acceptance of the wider research communities, incorporating incentives to push further frontiers of knowledge. It has to promote a culture of transparency supporting reproducibility of experiments for well-founded review. It should take into account established good practices and traditions which differ across scientific communities. The European Commission (EC) has been working on a framework of open science addressing in particular the impact from data, computing and networking infrastructures. Important steps were taken when launching Horizon 2020.

Open Science

As proposed in the "Open Science for the 21st century" declaration , open science can be unfolded in three components: open cultures, open content and open infrastructures. From the perspective of trust building, open science envisages optimal sharing of research data and also publications, software, and educational resources. The potential to mash-up, and to re-use research datasets will not only enable accurate scrutiny but will also reveal unexpected relationships and will trigger new findings. The European Commission is engaged to ensure an open access framework for publications stemming from EU-funded research and is progressively opening access to the research data (the basis for Horizon 2020). The EC is asking funding bodies in EU Member States to do the same.

Open infrastructures and the Research Data Alliance (RDA)

e-infrastructures are key components of the open science framework. They support advanced science and enable online research collaboration across disciplines at global level. They have the potential to structure the global knowledge space, increase scope, depth and economies of scale of the scientific enterprise. And, not least, they bridge the gap between scientists and the citizen and are enablers of trust in the scientific process. Data is a basic element of e-infrastructures. It has always been so but even more now at the dawn of "data-driven science" when e-infrastructures become a great opportunity for widening the participatory nature of science. The Riding the Wave and the Data Harvest reports highlight the strategic importance for Europe to support interoperability of research data infrastructures. They also point strongly at the need to support cost-effective research data management and the emergence of a computing literate generation of researchers in all fields of science. The European Commission is supporting the development of a pan-European multi-disciplinary data infrastructure through Horizon 2020 and policy developments centred on openness and interoperability. The global Research Data Alliance will support the EC strategy to achieve global scientific data interoperability in a way that real actors (users and producers of data, service providers, network and computing infrastructures, researchers and their organisations) are in the driving seat. Investments in digital infrastructure are needed to ensure that Europe remains a central hub for research and innovation, offering the best infrastructure to the brightest minds in the world.

Final remarks

Universality of science requires trusted and equitable access across all economic and social sectors. An open science framework will help fostering transparency and integrity and therefore trust in the scientific enterprise. An open science/open e-infrastructure framework should preserve the incentives of scientific discovery and the need to share and trust in order to collaborate across disciplinary and geographical boundaries, and also to develop the infrastructure capacity to support innovation. All stakeholders of the science, education and innovation "ecosystem" should promote practical applications of open science principles. For an open science framework to emerge the active contribution of many different players is necessary: from policy makers and funders to the individual researcher and ultimately to the engaged citizen. It requires a strong coordination effort at European and global levels and the promotion of global interoperability of data infrastructures through community-led initiatives such as the Research Data Alliance.

Carlos Morais Pires

Links:

Riding the Wave Report: http://kwz.me/Df
Data Harvest report: http://kwz.me/Dj
Research Data Alliance: https://rd-alliance.org/
Science as an open enterprise, The Royal Society Science Policy Centre, June 2012: http://kwz.me/Dq
Open Science Declaration: http://kwz.me/Dv

Introduction to the Special Theme

Scientific Data Sharing and Re-use

by Costantino Thanos and Andreas Rauber

Research data are essential to all scientific endeavours. Openness in the sharing of research results is one of the norms of modern science. The assumption behind this openness is that scientific progress requires results to be shared within the scientific community as early as possible in the discovery process.

The emerging cultures of data sharing and publication, open access to, and reuse of data are the positive signs of an evolving research environment. Data sharing and (re)usability are becoming distinct characteristics of modern scientific practice, as they allow reanalysis of evidence, reproduction and verification of results, minimizing duplication of effort, and building on the work of others. However, several challenges still prevent the research community from realizing the full benefits of these practices.

Data sharing/reusability has four main dimensions: policy, legal, technological and economic. A legal and policy framework should favour the open availability of scientific data and allow legal jurisdictional boundaries to be overcome, while technology should render physical and semantic barriers irrelevant. Finally, data sharing/reuse involves economic support: who will pay for public access to research data?

To make scientific data shareable and usable it should be discoverable, i.e. scholars must be able to quickly and accurately find data that supports scientific research; understandable to those scrutinizing them; and assessable enabling potential users to evaluate them. An emerging 'best practice' in the scientific method is the process of publishing scientific data. Data Publication refers to a process that allows a data user to discover, understand, and make assertions about the trustworthiness and fitness for purpose of the data. In addition, it should allow data creators to receive academic credit for their work.

The main technological impediments to data sharing/reuse are:

- *Heterogeneity of Data Representations:* There are a wide variety of scientific data models and formats and scientific information expressed in one formalism cannot directly be incorporated into another formalism.
- *Heterogeneity of Query Languages:* Data collections are managed by a variety of systems that support different query languages.
- *Discoverability of data:* In a networked scientific multidisciplinary environment pinpointing the location of relevant data is a big challenge for researchers.
- *Understandability of data:* The next problem regards the capacity of the data user to understand the information/knowledge embodied in it.
- *Movement of data:* Data users and data collections inhabit multiple contexts. The intended meaning becomes distorted when the data move across semantic boundaries. This is due to the loss of the interpretative context and can lead to a phenomenon called "ontological drift". This risk arises when a shared vocabulary and domain terminology are lacking.
- *Data Mismatching:* Several data mismatching problems hamper data reusability:
  - Quality mismatching occurs when the quality profile associated with a data set does not meet the quality expectations of the user of this data set.
  - Data-incomplete mismatching occurs when a data set is lacking some useful information (for example, provenance, contextual, uncertainty information) to enable a data user to fully exploit it.
  - Data abstraction mismatching occurs when the level of data abstraction (spatial, temporal, graphical, etc.) created by a data author does not meet the expected level of abstraction by the data user.

Standards

The role of standards in increasing data understandability and reusability is crucial. Standardization activities characterize the different phases of the scientific data life-cycle. Several activities aim at defining and developing standards to:

- represent scientific data - i.e., standard data models
- query data collections/databases - i.e., standard query languages
- model domain-specific metadata information - i.e., metadata standards
- identify data - i.e., data identification standards
- create a common understanding of a domain-specific data collection - i.e., standard domain-specific ontologies/ taxonomies and lexicons
- transfer data between domains - i.e., standard transportation protocols, etc.

A big effort has been devoted to creating metadata standards for different research communities. Given the plethora of standards that now exist, some attention should be directed to creating crosswalks or maps between the different standards.

This special issue features a keynote paper from an EU funding organization, an invited paper from a global organization that aims to accelerate and facilitate research data sharing and exchange, an invited paper from a prominent US scientist and an invited paper from a large Australian data organization. The core part of this issue presents several contributions of European researchers that address the different aspects of the data sharing and (re)use problem.

Please contact:
Costantino Thanos, ISTI-CNR, Italy,
E-mail: thanos@isti.cnr.it

Andreas Rauber, TU Vienna, Austria
E-mail: rauber@ifs.tuwien.ac.at

Creating the Culture and Technology for a Global Data Infrastructure

by Mark A. Parsons

The Research Data Alliance implements data sharing infrastructure by building social and technical bridges across cultures, scales and technologies.

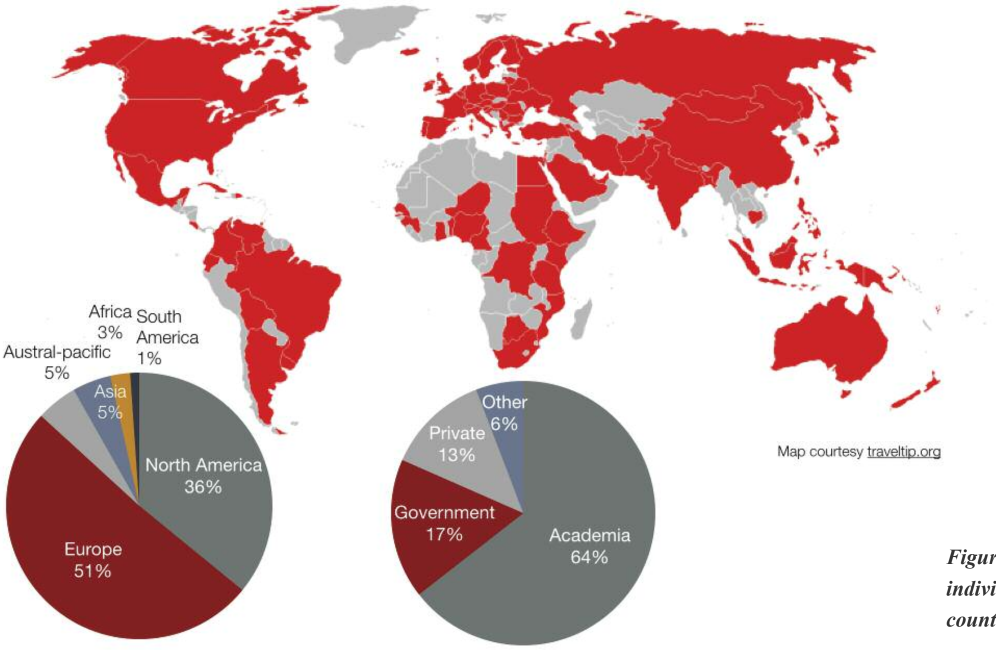
Research data are central to human understanding and a sustained, healthy society. Data provide the foundation for the evidence-based research that has advanced and transformed society in recent centuries. Data validate the theories that create the basis of human knowledge. Now we have bigger and more varied data than ever and huge societal challenges. Some say we have entered a new paradigm of research where data exploration is leading us to new theories and understanding rather than theories guiding our data collection [1]. This new paradigm will create new social and economic opportunities. Much like the Internet evolved from an academic network to a global communication infrastructure that fundamentally changed commerce and employment, ready and rapid data access can create new forms of wealth and transform society as well as research. This transformation can only occur if there is the culture and technology of a supporting and adaptive global data infrastructure.

The Research Data Alliance (RDA, rd-alliance.org) formed in 2013 to accelerate this transformation. RDA is a global, community-based, member organisation,working to build the social and technical bridges that enable open sharing of data.

The bridge is an important metaphor. Bridges are critical to infrastructure. So with data infrastructure we seek to bridge and share data across technologies, scales, disciplines, and cultures to address the grand challenges of society. But building infrastructure is incredibly complex. It is not just pipes and wires, but also relationships connecting machines, people and organisations. If we consider how past infrastructures developed, it is clear that infrastructure evolves somewhat organically. It is never really designed or architected at the outset [2].

We have seen time and again how top-down, “build-it-and-they-will-come” systems do not realize their potential or simply fail. RDA strives to be more bottom-up, allowing anyone to join the organization if they agree to our basic principles of openness, balance and harmonization through a community-driven, consensus-based, non-profit approach. Members can work on whatever problem is important to them as long as it advances data sharing. We’re not trying to solve all the data problems. We focus on implementing data sharing solutions. We aim to balance a grassroots approach with just enough guidance and process to succeed in implementing infrastructure.

RDA is also about people and the work they do. In less than two years, we have more than 2,500 members from Europe, the United States, Australia and many other countries. Members are mostly academics, but there is increasing representation from the government and private sectors (Figure 1). We also have about two-dozen Organisational Members including tech companies, libraries and regional efforts like the Barcelona Supercomputing Center and EUDAT. These organisations are key to ensuring the relevance and adoption of RDA products.



Map courtesy traveltip.org

Figure 1: Distribution of 2,538 individual RDA members in 92 countries as of 3 December 2014.

All these people and organisations come together in short-term, tiger-team-style Working Groups and broader, more exploratory Interest Groups. Currently, 14 Working Groups and 31 Interest Groups are addressing myriad topics. Some groups are technical—addressing issues such as persistent identifiers and machine actionable rules and workflows. Others address social issues such as legal interoperability or repository certification. Some bridge the social and technical on issues like data citation or best practices for repositories, while others are specific to certain disciplines or domains like agriculture, oceanography, genomics, and history and ethnography.

Working Groups only exist for 12-18 months. At the end of which, they deliver outputs, which could be a particular specification, method or practice that improves data sharing. To be approved as a Working Group, the group must demonstrate it has members who actually plan to use what is developed. This helps focus the work and ensures relevance. Many groups are co-sponsored by partner organisations, including ORCID, DataCite, the Data Seal of Approval, CODATA, and the World Data System.

RDA members work at multiple scales from the global to the local. Implementation of outputs is inherently a local activity, but it is most relevant and impactful if it is done in a global context. So, in addition to the global RDA, there are local or regional RDAs. RDA/Europe includes all the European members in RDA and works to ensure that RDA is relevant to unique European needs. See for example the recent RDA/Europe report, The Data Harvest: How sharing research data can yield knowledge, jobs and growth [3].

In short, RDA is about implementing data infrastructure by people working together at all scales. It is an exciting opportunity, and it highlights the power of volunteer effort towards addressing grand challenges. RDA is clearly relevant to members of ERCIM who are encouraged to join RDA as Organisational Members and attend our next Plenary 9-11 March in San Diego.

Acknowledgements: Yolanda Meleco, RDA/US, provided very useful edits to this article.

Link: https://rd-alliance.org/

References:

[1] T. Hey, S. Tansley, and K. Tolle (eds): The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009

[2] P.N. Edwards, S.J. Jackson, G.C. Bowker, C.P. Knobel: "Understanding Infrastructure: Dynamics, Tensions, and Design", NSF. 2007 http://hdl.handle.net/2027.42/49353

[3] F. Genova, H. Hanahoe, L. Laaksonen, C. Morais-Pires, P. Wittenburg and J. Wood: "The Data Harvest: How sharing research data can yield knowledge, jobs and growth", RDA Europe, 2014.

Please contact:

Mark Parsons, Secretary General of the Research Data Alliance (RDA)
Rensselaer Polytechnic Institute, USA
E-mail: parsom3@rpi.edu

If Data Sharing is the Answer, What is the Question?

by Christine L. Borgman

Data sharing has become policy enforced by governments, funding agencies, journals, and other stakeholders. Arguments in favor include leveraging investments in research, reducing the need to collect new data, addressing new research questions by reusing or combining extant data, and reproducing research, which would lead to greater accountability, transparency, and less fraud. Arguments against data sharing rarely are expressed in public fora, so popular is the idea. Much of the scholarship on data practices attempts to understand the socio-technical barriers to sharing, with goals to design infrastructures, policies, and cultural interventions that will overcome these barriers.

However, data sharing and reuse are common practice in only a few fields. Astronomy and genomics in the sciences, survey research in the social sciences, and archaeology in the humanities are the typical exemplars, which remain the exceptions rather than the rule. The lack of success of data sharing policies, despite accelerating enforcement over the last decade, indicates the need not just for a much deeper understanding of the roles of data in contemporary science but also for developing new models of scientific practice. Science progressed for centuries without data sharing policies. Why is data sharing deemed so important to scientific progress now? How might scientific practice be different if these policies were in place several generations ago?

Enthusiasm for "big data" and for data sharing are obscuring the complexity of data in scholarship and the challenges for stewardship [1]. Data practices are local, varying from field to field, individual to individual, and country to country. Studying data is a means to observe how rapidly the landscape of scholarly work in the sciences, social sciences, and the humanities is changing. Inside the black box of data is a plethora of research, technology, and policy issues. Data are best understood as representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship. Rarely do they stand alone, separable from software, protocols, lab and field conditions, and other context. The lack of agreement on what constitutes data underlies the difficulties in sharing, releasing, or reusing research data.

Concerns for data sharing and open access raise broader questions about what data to keep, what to share, when, how, and with whom. Open data is sometimes viewed simply as releasing data without payment of fees. In research contexts, open data may pose complex issues of licensing, ownership, responsibility, standards, interoperability, and legal harmonization. To scholars, data can be assets, liabilities, or both. Data have utilitarian value as evidence, but they also serve social and symbolic purposes for control, barter, credit, and prestige. Incentives for scientific advancement often run counter to those for sharing data.

To librarians and archivists, data are scholarly products to curate for future users. However, data are more difficult to manage than publications and most other kinds of evidence. Rarely are data self-describing, and rarely can they be interpreted outside their original context without extensive documentation. Interpreting scientific data often requires access to papers, protocols, analytical tools, instruments, software, workflows, and other components of research practice – and access to the people with whom those data originated. Sharing data may have little practical benefit if the associated hardware, software, protocols, and other technologies are proprietary, unavailable, or obsolete and if the people associated with the origins of the data cannot be consulted [2, 3].

Claims that data and publications deserve equal status in scholarly communication for the purposes of citation raise a host of theoretical, methodological, and practical problems for bibliometrics. For example, what unit should be cited, how, when, and why? As argued in depth elsewhere, data are not publications [1]. The "data publication" metaphor, commonly used in promoting open access to data and encouraging data citation, similarly muddies the waters. Transferring bibliographic citation principles to data must be done carefully and selectively, lest the problems associated with citation practice be exacerbated and new ones introduced. Determining how to cite data is a non-trivial matter.

Rather than assume that data sharing is almost always a "good thing" and that doing so will promote the progress of science, more critical questions should be asked: What are the data? What is the utility of sharing or releasing data, and to whom? Who invests the resources in releasing those data and in making them useful to others? When, how, why, and how often are those data reused? Who benefits from what kinds of data transfer, when, and how? What resources must potential re-users invest in discovering, interpreting, processing, and analyzing data to make them reusable? Which data are most important to release, when, by what criteria, to whom, and why? What investments must be made in knowledge infrastructures, including people, institutions, technologies, and repositories, to sustain access to data that are released? Who will make those investments, and for whose benefit?

Only when these questions are addressed by scientists, scholars, data professionals, librarians, archivists, funding agencies, repositories, publishers, policy makers, and other stakeholders in research will satisfactory answers arise to the problems of data sharing [1].

References:

[1] C.L. Borgman: "Big Data, Little Data, No Data: Scholarship in the Networked World". MIT Press, 2015.

[2] C.L. Borgman et al.: "The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management", ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) and International Conference on Theory and Practice in Digital Libraries (TPDL 2014) (London, 2014), 2014.

[3] J.C. Wallis et al.: "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology", PLoS ONE. 8, 7 (Jul. 2013), e67332.

Please contact:

Christine L. Borgman
University of California, Los Angeles, USA
E-mail: Christine.Borgman@ucla.edu

Enhancing the Value of Research Data in Australia

by Andrew Treloar, Ross Wilkinson, and the ANDS team

Over the last seven years, Australia has had a strong investment in research infrastructure, and data infrastructure is a core part of that investment.

Much has been achieved already. The Government understands the importance of data, our research institutions are putting in place research data infrastructure, we can store data, we can compute over data, and our data providing partners – research institutions, public providers, and NCRIS data intensive investments are ensuring that we are establishing world best data and data infrastructure.

The Australian National Data Service (ANDS) commenced in 2009 to establish an Australian research data commons. It has progressively refined its mission towards making data more valuable to researchers, research institutions and the nation. Over the last 5 years ANDS has worked across the whole sector in partnership with major research organisations and NCRIS facilities. It has worked collaboratively to make data more valuable through bringing about some critical data transformations: moving to structured data collections that are managed, connected, discoverable and reusable. This requires both technical infrastructure and community capability, and can deliver significant research changes [1].

We have seen many examples where these transformations have been suc-

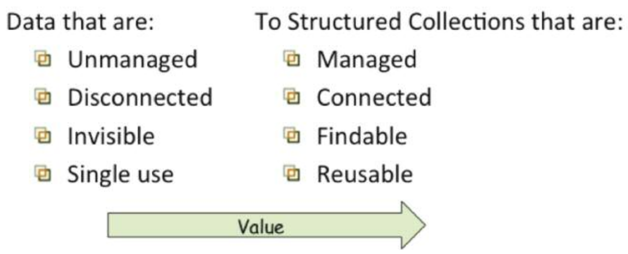


Figure 1: Transformation of Data.

cessful. We give three examples, showing the technical, community and the research effects.

Parkes Observatory Pulsar Data Archive

Information gathered over many years using CSIRO's Parkes Radio Telescope is being used as a major resource in the international search for gravitational waves. This is one of the more unexpected outcomes of the construction of the Parkes Observatory Pulsar Data Archive [2]. This project fulfilled a CSIRO commitment to the world's astronomers - to make data from the Parkes telescope available publicly within 18 months of observation.

The data archive was established with support from the ANDS. It also has freed CSIRO astronomers from the time consuming task of satisfying requests for the Parkes data from all over the world. Those requests come flooding in because Parkes is where the bulk of all known pulsating neutron stars or pulsars have been discovered.

In addition, astronomers from Peking University, who are using the Parkes data archive as a training tool for radio telescope data analysis, have published several papers which include descriptions of pulsars and other astronomical bodies they have newly discovered.

The archive is accessible through both Research Data Australia (the national data portal) and CSIRO's Data Access Portal.

Data Citation Support

Citation is a pivotal concern in the publication and reuse of data. ANDS' approach to raising awareness has resulted in a vibrant and burgeoning Australian community now routinely applying DOIs to data. Community were engaged through webinars, roundtables, workshops, train-the- trainer sessions, seminars YouTube recordings, and countless resource links were provided, including links to key international resources.

As well as community development there was corresponding technical support through the ANDS Cite My Data Service. Technical documentation is accompanied by plain–English information to help institutions effect cultural change. As well many additional resources supporting the community were made available through the ANDS website.

As a result 26 institutions in Australia are minting DOIs (Digital Object Identifiers) through the ANDS CiteMyData service and citation metrics are being harvested by data collecting agencies. The growing practice of data citation is underpinned and driven forward by librarians who are knowledgeable and skilled in research data management and reuse, and by institutionally focused services and materials.

The Research Student

A PhD student is quietly changing the way we assess our fisheries without setting foot on a boat. By analysing more than 25 years of Australian Fisheries Management Authority records, he has found that a key assumption of the models employed to make predictions of the future of Australia's ocean fish stocks is not justified [3]. "It has always been assumed for each different species that the relationship of length with age does not change through time," says Mr Athol Whitten from CSIRO and the University of Melbourne. By going back and pulling the data apart, Whitten found that some species are showing a long-term decline of length with age, and in others, the growth rate depends on the year they were born. The information now amounts to more than 180,000 data points on 15 species of fish over 25 years. Access to this wealth of information has provided Whitten with an efficient way to pursue his research on testing the assumptions of the fisheries models.

In Summary

There is now a strong research data management capacity in Australia that uses a coherent approach to Australia's research data assets, and can support substantial change research data practice in the light of policy or technical changes.

Significant progress has been made in enabling improved data management, connectivity, discoverability, and usability by:

- Establishing the Australian Research Data Commons, a network of shared data resources
- Populating the Australian Research Data Commons with over 100,000 research data collections
- Dramatically improving institutional research data management capacity
- Helping to establish institutional research data infrastructure
- Co-leading the establishment of the Research Data Alliance, improving international data exchange.

This has meant that Australian researchers, research institutions and the nation are at the forefront of the opportunities inherent in global research data intensive activity.

Link:
http://www.ands.org.au

References:

[1] Material drawn from Australian National Data Service, http://ands.org.au/, Accessed, 2014
[2] G. Hobbs et al.: "The Parkes Observatory Pulsar Data Archive", Publications of the Astronomical Society of Australia 28, 202, 2011
[3] A. Whitten, N. Klaer, G. Tuck, R. Day: "Variable Growth in South-Eastern Australian Fish Species: Evidence and Implications for Stock Assessment", 141st American Fisheries Society Annual Meeting held in Seattle, 2011

Please contact:

Ross Wilkinson
Australian National Data Service
E-mail: ross.wilkinson@ands.org.au

Beyond Data: Process Sharing and Reuse

by Tomasz Miksa and Andreas Rauber

Sharing and reuse of data is just an intermediate step on the way to reproducible computational science. The next step, sharing and reuse of processes that transform data, is enabled by process management plans, which benefit multiple stakeholders at all stages of research.

Nowadays almost every research domain depends on data that is accessed and processed using computers. Data processing may range from simple calculations made in a spreadsheet editor, to distributed processes that transform data using dedicated software and hardware tools. The crucial focus is on the data, because it underlies new scientific breakthroughs and in many cases is irreproducible, e.g. climate data. Funding institutions have recognized the value of data, and as a result data management plans (DMPs) have become obligatory for many scientists who receive public funding. Data management plans are initiated before the start of a project, and evolve during its course. They aim not only to ensure that the data is managed properly during the project, e.g. by performing backups, using file naming convention, etc., but also that it is preserved and available in the future.

In order to understand data, as well as research results, data acquisition and manipulation processes must also be curated. Unfortunately, the underlying processes are not included in DMPs. As a consequence, information needed to document, verify, preserve or re-execute the experiment is lost. For this reason, we extend DMPs to “process management plans” (PMPs) [1] which complement the description of scientific data taking a process-centric view, viewing data as the result of underlying processes such as capture, (pre-) processing, transformation, integration and analyses. A PMP is a living document, which is created at the time of process design and is maintained and updated during the lifetime of the experiment by various stakeholders. Its structure as outlined below necessarily appears relatively abstract due to a wide range of requirements and different practices in scientific domains in which PMPs should be used. The proposed structure of a PMP is depicted in Figure 1.

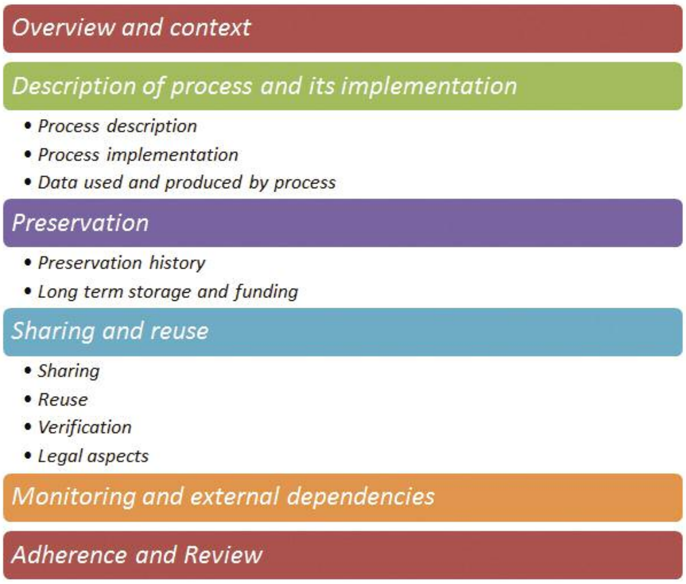


Figure 1: Structure of a Process Management Plan.

The implementation of PMPs is to some extent domain dependent, because it has to incorporate already existing best practices. During the course of the EU-funded FP7 project TIMBUS, our team at SBA Research in Vienna investigated well-structured Taverna workflows, but also unstructured processes from the civil-engineering domain. In all cases, the PMPs can be implemented by integrating already existing tools. For example, the automatically generated TIMBUS Context Model [2] can be used to describe the software and hardware dependencies. The process of verification and validation can be performed using the VFramework. Research Objects [3] can be used to aggregate the high level information on the process, and the existing data management plan templates and tools can be refined to incorporate information on processes.

Figure 2 depicts stakeholders impacted by PMPs. Project applicants will benefit by being able to better identify and plan the resources needed for the research. For example, if the process of transforming the experimental data assumes use of proprietary software with an expensive license, this information will be revealed at an early stage and can be taken into the account when applying for a grant.

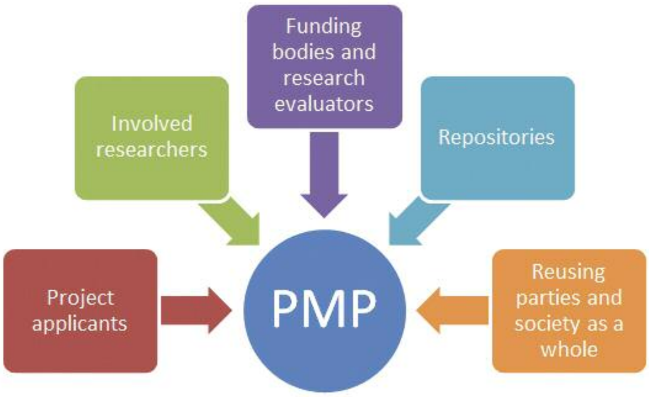

Funding bodies and research evaluators
Involved researchers
Repositories
Project applicants
PMP
Reusing parties and society as a whole

Figure 2: Stakeholders impacted by Process Management Plan.

Researchers will benefit by working with better documented processes. This leverages sharing of results and eases reuse of existing research. Moreover, time will be saved when a researcher joins a new project, because useful information will be provided in a structured way by the PMP. This is especially important for cooperation within research infrastructures, where research challenges are addressed through cooperation of various specialists from different areas contributing to only a specific part of the experiment.

From the point of view of funding bodies, PMPs safeguard the investment made into research by ensuring research results are trustable /verifiable, and can be re-used at later points in time. Furthermore, PMPs facilitate cooperation between projects, because they make it easier to reuse processes used in other projects and facilitate exploitation of results of other funded projects. Thus, PMPs lead to sustainable research and can save funding that can be allocated to new research projects.

Repositories which keep the deposited processes and data can better estimate the costs of curation and plan actions needed to maintain the deposited resources. PMPs also support long term preservation (keeping process usable over time) and provide information on possible events triggering necessary digital preservation activities.

PMPs also benefit reusing parties, because their research can be accelerated by reusable processes. The reusing parties also have greater confidence that they can build on previous work, because the quality is higher due to the reproducibility. Furthermore, scientists whose processes are reused for other experiments gain recognition and credit.

We are currently working on automation of PMP creation and verification by extraction of process characteristics automatically from its environment. Specific focus is on tool support to automate many of the various documentation steps. We are also currently evaluating the PMP with stakeholders from different scientific communities.

Links:
TIMBUS project:
http://timbusproject.net
SBA Research: http://www.sba-research.org/
Process description tools:
http://www.ifs.tuwien.ac.at/dp/process/
DMP tools:
http://www.dcc.ac.uk/dmponline

References:

[1] Miksa et al.: “Process Management Plans”, International Journal of Digital Curation.
[2] Mayer et al.: “Preserving scientific processes from design to publication”, in proc. of TPDL 2012.
[3] Bechhofer et al.: “Research Objects: Towards Exchange and Reuse of Digital Knowledge”, Nature Precedings.

Please contact:

Tomasz Miksa, SBA, Austria
Tel: +43 69915089237
E-mail: tmiksa@sba-research.org

Andreas Rauber, TU Vienna, Austria
Tel: +43 15880118899
E-mail: rauber@ifs.tuwien.ac.at

Open Data – What do Research Communities Really Think about it?

by Marie Sandberg, Rob Baxter, Damien Lecarpentier and Paweł Kamocki

Facilitating open access to research data is a principle endorsed by an increasing number of countries and international organizations, and one of the priorities flagged in the European Commission’s Horizon 2020 funding framework [1][2]. But what do researchers themselves think about it? How do they perceive the increasing demand for open access and what are they doing about it? What problems do they face, and what sort of help are they looking for?

As a pan-European research data infrastructure, these are questions that are of fundamental interest to EUDAT. To better understand what researchers think, EUDAT has conducted a programme of interviews with fourteen major research communities from the fields of life sciences, Earth and atmospheric science, astrophysics, climate science, biodiversity, agricultural science, social science and humanities – a broad cross-section of European research interests. While the views of any given individual cannot be interpreted as the official position of a whole research community, they nevertheless provide useful information about the general attitude, requirements and challenges researchers face with regard to opening up their research data. In this article we report our initial conclusions from this survey.

Growing Awareness

Open access to research data is increasingly seen as a compelling principle in many research communities. There is a growing awareness of the global move towards open access, the potential benefits it may offer, and the need to implement open access policies within particular disciplines. According to preliminary figures on the first wave of open data pilot projects in Horizon 2020, the opt-out rate among proposals submitted to the “open by default” categories was below 30%, and the opt-in rate among other proposals was around about the same. This underlines our findings in EUDAT – researchers are pretty happy about sharing their data.

Challenges Ahead

In practice, though, there are many unsolved challenges still to be addressed, and those most often cited by researchers were the ethical and legal complications, and the issue of credit.

Not all data can be made open access. Personal data, and especially sensitive personal data, is particular challenging. In these days of large-scale combination and data mining, can such data truly be anonymized for research purposes [3]? And what about the re-purposing of data for ends very far away from the original research agenda – for military or even criminal purposes? There are no easy answers to these questions, and the culture of ethics surrounding good research is making some communities tread warily.

Our survey highlights a lack of knowledge about the legal aspects of data sharing and data reuse, in particular around intellectual property rights, copyright and licensing, which can act as a barrier not only for opening data but also for re-using someone else's data. Choosing the right licence, for instance, can be a daunting task for some researchers who don't necessarily understand the implications of their actions.

While researchers are naturally keen to see their research published as widely as possible, in an interesting contrast to the open access scholarly paper movement, open data is viewed differently. Often research groups invest significant time and effort in collecting "hard to get data" which can then be used to build careers, offering what can only be termed a competitive advantage over those who do not have access to the same data. This issue of credit and consequent career progression is a real concern in many communities.

The way forward

While aware of, and supportive of, the open access data agenda, many research communities are looking for guidance about the practicalities of doing it; training on managing the legal issues, for instance. They also feel that these issues should be addressed at cross-disciplinary level, perhaps rendering the tasks even more challenging. And while much of the open access focus is on coordination efforts, training needs and policies, researchers also stress the importance of developing the right tools and services to enable these policies and, ultimately, the sharing and reuse of data; this is seen as particular crucial for handling sensitive data.

Some final words

Compared to scholarly publications, open access to research data is both less developed and more difficult to implement. Although open access to research data has only just begun, the broad spectra of expectations on EUDAT and other initiatives show that research communities have the notion that open access to research data cannot be solved through isolated activities or actions; instead it needs to underpin the whole system, reaching from strategic planning and overall polices to the mindset and everyday practice of the individual researcher.

Link:
EUDAT – European Data project:
http://www.eudat.eu/

References:

[1] European Commission: "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020".
[2] G8 Science Ministers Statement, 13 June 2013, available at https://www.gov.uk/government/news/g8-science-ministers-statement
[3] Article 29 Data Protection Working Party: "Opinion 05/2014 on Anonymisation Techniques", adopted on 10 April 2014.

Please contact:

Marie Sandberg
CSC, Finland
Tel: +358 9 457 2001
E-mail: marie.sandberg@csc.fi

Rob Baxter
University of Edinburgh, UK
Tel: +44 131 651 3579
E-mail: r.baxter@epcc.ed.ac.uk

Providing Research Infrastructures with Data Publishing

by Massimiliano Assante, Leonardo Candela, Paolo Manghi, Pasquale Pagano, and Donatella Castelli

The purpose of data publishing is to release research data for others to use. However, its implementation remains an open issue. 'Science 2.0 Repositories' (SciRepos) address the publishing requirements arising in Science 2.0 by blurring the distinction between research life-cycle and research publishing. SciRepos interface with the ICT services of research infrastructures to intercept and publish research products while providing researchers with social networking tools for discovery, notification, sharing, discussion, and assessment of research products.

Data publishing approaches, namely the "data as a paper" ones [1], are mainly inspired by scientific literature communication workflows, which separate the place where research is conducted, i.e., Research Infrastructures (RIs), from the place where research is published and shared. In particular, research products are published "elsewhere" and "on date", i.e. when the scientists feel the products obtained so far are sufficiently mature. In our opinion, this model does not fit well with other kinds of research products, for which effective interpretation, evaluation, and reuse can be ensured only if publishing includes the properties of "within" the RIs and "during" the research activity.

To enable effective scientific communication workflows, research product creation and publishing should both occur "within" the RI (as opposed to "elsewhere") and "during" the research activities (as opposed to "on date"). To facilitate this, research infrastructure ICT services should not only be devised to provide scientists with facilities for carrying out their research activities, but also to support marketplace like facilities, enabling RI scientists to publish products created by research activities and other scientists to discover and reuse them. In other words, RIs should not rely on third-party marketplace sources to publish their products, but rather should integrate them into the RI.

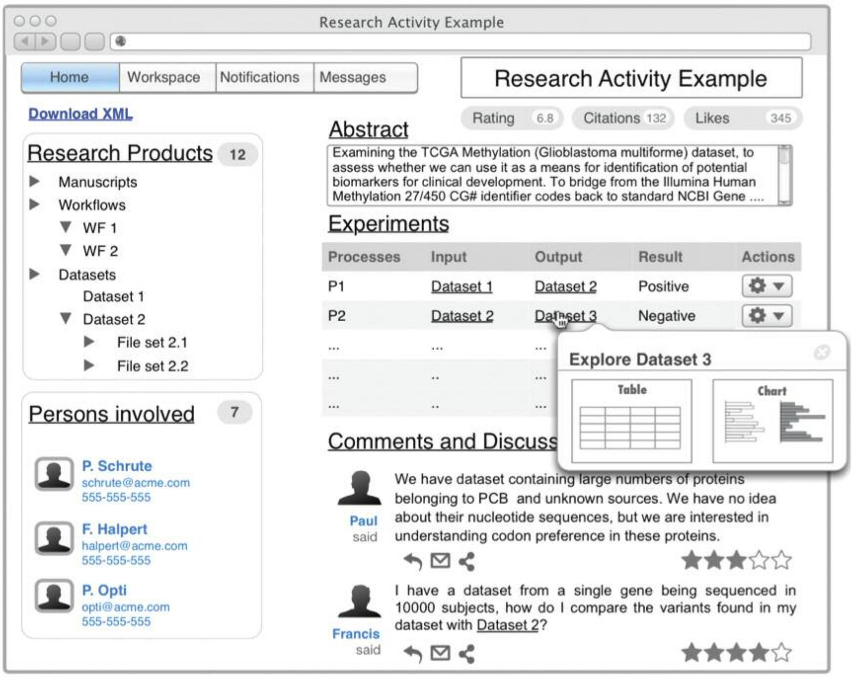
Unfortunately, current repository platforms are not suitable to implement this vision, as they are designed not to integrate with existing RI ICT services but instead to support today's notion of the "elsewhere" and "on date" research marketplace. We propose an innovative class of repositories: Science 2.0 Repositories (SciRepos).

SciRepos are characterized by the following features:

- Integration with RI ICT services in order to intercept the generation of products within research activities and to publish such products, i.e. making them discoverable and accessible to other researchers;
- Provision of repository-like tools so that scientists can access and share research products generated during their research activities;
- Dependence on social networking practices in order to modernize (scientific) communication both intra-RI and inter-RI, e.g., posting rather than deposition, "like" and "open discussions" for quality assessment, sharing rather than dissemination.

The SciRepo supports scientists with two kinds of end-user functionalities:

- Repository-oriented facilities: offering typical repository facilities on the information graph such as search and browse allowing search by product typology, but also permitting navigation from research activities to products and related products. Ingestion facilities are provided, allowing scientists to manually or semi-automatically upload "external" products into the repository and associate them with a research activity, thus including them in the information graph. Examples are publications, but also alternative scientific products, such as web sites, blogs, slides, documentation, manuals, etc. Ingestion allows scientists to complete the action of publishing a research activity with all products that are connected to it but generated out of the boundaries of the RI. The way scientists or groups of scientists can interact with products (access and reuse them) is ruled by clear rights management functionalities. Rights are typically assigned when products are generated in the RI or ingested by scientists, but can vary overtime.



Examining the TCGA Methylation (Glioblastoma multiforme) dataset, to assess whether we can use it as a means for identification of potential biomarkers for clinical development. To bridge from the Illumina Human Methylation 27/450 CG# identifier codes back to standard NCBI Gene ....

Processes	Input	Output	Result	Actions
P1	Dataset 1	Dataset 2	Positive	⚙ ▾
P2	Dataset 2	Dataset 3	Negative	⚙ ▾
...	...	...		
...	..	...		
...	..	...		

We have dataset containing large numbers of proteins belonging to PCB and unknown sources. We have no idea about their nucleotide sequences, but we are interested in understanding codon preference in these proteins.

★★★☆☆

I have a dataset from a single gene being sequenced in 10000 subjects, how do I compare the variants found in my dataset with Dataset 2?

★★★★☆

Figure 1: Repo: An example of a SciRepo Activity Web Page.

- Collaboration-oriented facilities: offering typical social networking facilities such as the option to subscribe to events that are relevant to research activities and products, and be promptly notified, e.g., the completion of a workflow execution, the generation of datasets obeying to some criteria. Users can reply to posts and, most importantly, can express opinions on the quality of products, e.g., "like" actions or similar. This goes in the direction of truly "open" peer-review. More sophisticated assessment/peer-review functionalities (single/double blind) can be supported, in order to provide more traditional notions of quality. Interestingly, the posts themselves represent a special typology of products of the research activity and are searchable and browsable in the information graph.

In order to implement a SciRepo, RIs should develop their own software, thereby investing in a direction that requires different kinds of skills and dedicated funds. To facilitate this process we are designing and developing a SciRepo platform, devised to support the implementation of SciRepos at minimum development cost for the RIs. This platform builds upon previous works and experience [2][3].

References:

[1] L. Candela et al. "Data Journals: A Survey", Journal of the Association for Information Science Science and Technology, 2014.

[2] M. Assante et al.: "A Social Networking Research Environment for Scientific Data Sharing: The D4Science Offering", The Grey Journal, Vol. 10, 2014.

[3] A. Bardi and P. Manghi: "A rationale for enhanced publications", LIBER Quarterly, 2014.

Please contact:

Massimiliano Assante
ISTI-CNR, Italy
E-mail:
massimiliano.assante@isti.cnr.it

Sailing Towards Open Marine Data: the RITMARE Data Policy

by Anna Basoni, Stefano Menegon and Alessandro Sarretta

A thorough understanding of marine and ocean phenomena calls for synergic multidisciplinary data provision. Unfortunately, much scientific data is still kept in drawers, and in many cases scientists and stakeholders are unaware of its existence. At the same time, researchers lament the time consuming nature of data collection and delivery. To overcome barriers to data access, the RITMARE project issued a data policy document, an agreement among participants on how to share the data and products either generated by the project activities or derived from previous activities, with the aim of recognizing the effort involved.

The RITMARE Flagship Project is one of the National Research Programs funded by the Italian Ministry for Education, University and Research, coordinated by the National Research Council (CNR) and involving the whole Italian marine research community. The project's main goal is the interdisciplinary integration of national marine research.

The objective of the RITMARE Data Policy (RDP) is to define common rules on how to access, use and share the information provided by the project for an improved understanding of the marine environment.

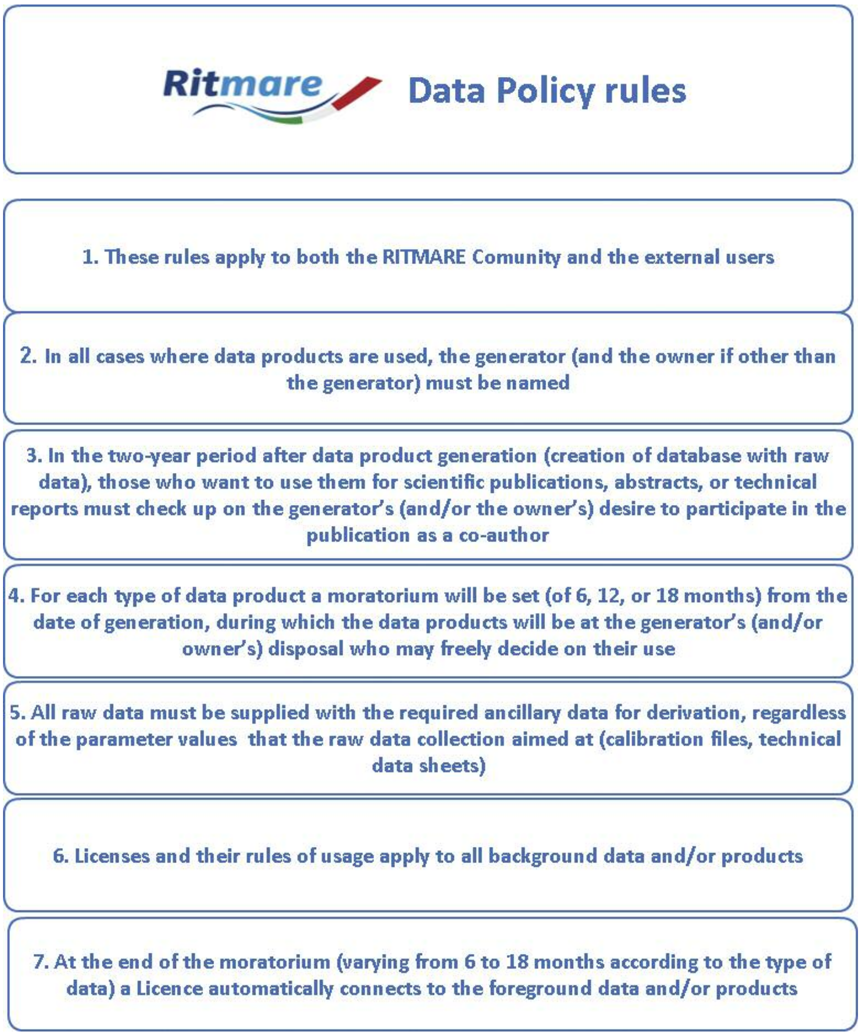
The RDP was formulated taking into account:

- the international scientific community, which strongly encourages the rapid sharing of information;
- the European Union which, in numerous directives, has clearly stated that free access to data and products generated by public funds is the only way to go [1];
- the Italian Digital Agenda, which is gradually formulating rules to govern open access [2].

Briefly, in accordance with several initiatives within the marine scientific context, the RDP tries to reach a trade-off between two conflicting needs: (i) the demand for easy and open access to data, and (ii) the requirements of data generators to see their work recognized and to have sufficient time to communicate their results and hypotheses.

The RDP does not apply to sensitive data, which is subject to agreements with protection and security institutions. Some other exemptions (potentially sensitive data for socio-economic reasons) can be evaluated by the RDP Governance Board. These rules represent the maximum level of data constraint accepted within the RITMARE project. Researchers, however, are encouraged to attribute less restrictive norms. In fact, even though licences that restrict the re-use of data to non-commercial purposes (e.g., the Creative Commons license CC-BY-NC) were accepted, the RDP recommends the use of less binding licences (e.g., CC-BY).

The highlights of the RDP are summarized in Figure 1, which illustrates the seven mandatory rules for the project. The first rule defines the involved parties: the entire RITMARE marine research community, both internal and external to the project. The second rule states that each user shall name the data generator; the third rule requires that everyone interested in publishing the data in scientific publications, abstracts and technical reports, within the first



Ritmare Data Policy rules

1. These rules apply to both the RITMARE Comunity and the external users

2. In all cases where data products are used, the generator (and the owner if other than the generator) must be named

3. In the two-year period after data product generation (creation of database with raw data), those who want to use them for scientific publications, abstracts, or technical reports must check up on the generator's (and/or the owner's) desire to participate in the publication as a co-author

4. For each type of data product a moratorium will be set (of 6, 12, or 18 months) from the date of generation, during which the data products will be at the generator's (and/or owner's) disposal who may freely decide on their use

5. All raw data must be supplied with the required ancillary data for derivation, regardless of the parameter values that the raw data collection aimed at (calibration files, technical data sheets)

6. Licenses and their rules of usage apply to all background data and/or products

7. At the end of the moratorium (varying from 6 to 18 months according to the type of data) a Licence automatically connects to the foreground data and/or products

Figure 1: RITMARE Data Policy rules.

two years from data creation, should verify the willingness of the data generator to participate as co-author. The fourth rule establishes the moratorium period: this could last from six to 18 months, depending on the data or product typology, during which period the generator has primary publication rights. After the moratorium period, a standard open licence has to be associated with the data, depending on the data type. To enable this automatic mechanism of licence attribution, a classification of data and products is provided by the generators (see seventh rule). Furthermore, to derive the correct data from the raw data, the fifth rule states that ancillary information should be available, containing, for example, the calibration file and instrument specifications.

The above rules refer to foreground data. For background data, the relevant licences or rules should be applied (see sixth rule). A data policy tool is available by means of which the appropriate licence will be assigned to each data/product. Every research group (generator) must fill in the table with the following data (metadata):

- owner of/responsible for each data/product
- name of data/product
- short description
- period of moratorium (in months)
- type of licence
- additional notes.

The RDP application contributes effectively to the full integration of the Italian marine scientific community through the transfer and sharing of data and knowledge in order to improve the quality of research produced, its visibility and its international competitiveness. Through integration with the other activities of the project, the RDP will help to make RITMARE a demonstrator of the importance of open access to information - not only in scientific research, but also to innovation and general quality of life.

The latest version of the document is now available on the Figshare repository [3], licensed under a Creative Commons Attribution 4.0.

Acknowledgement:
The activities described in this paper have been funded by the Italian Flagship Project RITMARE. The authors wish to thank Paola Carrara, Raffaele d'Adamo, Giuseppe Manzella, Gianfranco Pazienza, Anna Rampini and Maurizio Ribera d'Alcalà for their contribution and support.

Links:
RITMARE Flagship Project:
http://www.ritmare.it
Open definition:
http://opendefinition.org/
http://creativecommons.org/licenses/
http://www.dati.gov.it/iodl/2.0/
http://opendatacommons.org/licenses/odbl/
http://creativecommons.org/licenses/

References:
[1] European Commission: 2013/37/UE Directive of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information, 2013.
[2] Italian Law Decree n. 81/2013, in particular Art. 4.
[3] SP7_WP3, RITMARE Data Policy, http://figshare.com/articles/RITMARE_Data_Policy_document/1235546

Please contact:
Anna Basoni, IREA-CNR, Italy
E-mail: basoni.a@irea.cnr.it

Stefano Menegon, Alessandro Sarretta
ISMAR-CNR, Italy
E-mail: Stefano.Menegon|Alessandro
Sarretta@ismar.cnr.it

RDA: The Importance of Metadata

by Keith G. Jeffery and Rebecca Koskela

RDA is all about facilitating researchers to use data (including scholarly publications and grey literature used as data). This encompasses data collection, data validation, data management (including preservation/curation), data analysis, data simulation/modelling, data mining, data visualisation and interoperation of data. Metadata are the key to all of these activities because they present to persons, organisations, computer systems and research equipment a representation of the dataset so that the dataset can be acted upon.

Metadata are defined by some as ‘data about data’. In fact there is no difference between metadata and data except for the purpose for which they are used. An electronic library catalogue card system provides metadata for the researcher finding an article or book but data for the librarian counting the articles on biochemistry. Metadata are used both by humans and by computers; however for scalability and virtualisation (hiding unnecessary complexity from the end-user), it is necessary to ensure that metadata can be both read and ‘understood’ by computer systems. This leads to the mantra ‘formal syntax and defined semantics’: humans can overcome inconsistencies and vagueness but computers cannot.

Metadata Purposes

Metadata are used commonly for (a) discovery, (b) contextualisation and (c) detailed processing [1]. In the case of discovery the metadata must be sufficient for the human or computer system to find the datasets / data objects of interest for the purpose. The higher the quality of the discovery metadata, the greater the precision (accuracy) and recall (completeness) of the discovery. Typical ‘standards’ in the discovery area are DC (Dublin Core) and CKAN (Comprehensive Knowledge Archive Network)

In the area of contextual metadata perhaps the most widely used ‘standard’ (An EU Recommendation to Member States, used in 43 countries, adopted by Elsevier and Thomson-Reuters) is CERIF (Common European Research

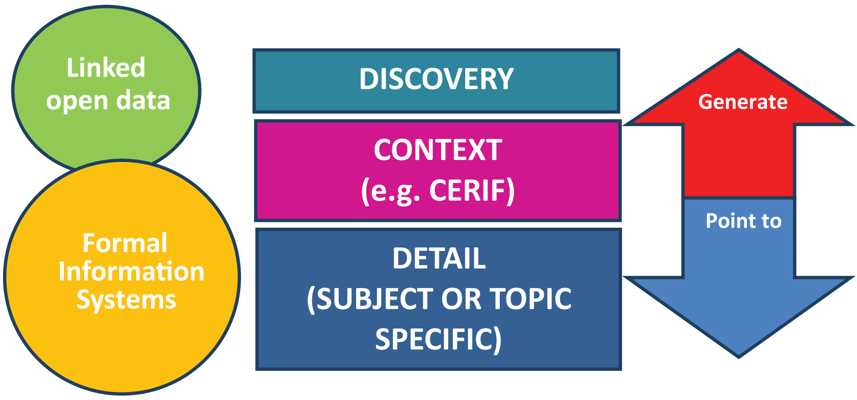


Figure 1: Purposes of Metadata and their Relationships

Information Format) which covers persons, organisations, projects, products (including datasets), publications, patents, facilities, equipment, funding and – most importantly – the relationships between them expressed in a form of first order logic with both role and temporal attributes [2].

The detailed processing metadata are typically specific to a research domain or even an individual experiment or observation method. They include schema information to connect software to data and also parameters necessary for correct data processing such as precision, accuracy or calibration information.

Metadata in RDA: As indicated above, metadata are used extensively in all aspects of RDA activity. However there are four groups that are specialising in metadata. They are: MIG (Metadata Interest Group): the overarching long-term group to discuss metadata and to work with Working Groups (WGs) of 18-month duration doing specific tasks; MSDWG (Metadata Standards Directory WG): developing a directory of metadata standards so a user can look up appropriate standards for their purpose and/or research domain; DICIG (Data in Context IG): developing through use cases the requirements within and across research domains for contextual metadata; RDPIG (Research Data Provenance IG): concentrating on providing provenance information for datasets. These groups arose spontaneously 'bottom-up' but are now coordinating among themselves to form a strong metadata presence in RDA.

Moving Forward

The metadata groups have agreed on a joint forward plan. It consists of the following steps:

1. Collect use cases: a form has been prepared and is available on the website together with a use case example both written and on the form;
2. Collect metadata 'standards' into the MSDWG directory;
3. Analyse content of (1) and (2) to produce a superset list of all elements required and a subset list of common elements by purpose – so called 'packages' of metadata elements:
4. Test those 'packages' with research domain groups in RDA (we have already volunteers!) and adjust based on feedback;
5. Present the 'packages' to the TAB (Technical Advisory Board) of RDA for authorising as recommendations from RDA to the community.

The metadata groups plan to meet jointly, and jointly with chairs of other groups, at RDA Plenary 5 in San Diego. You are welcome to join RDA, register for P5 and be involved.

Acknowledgements

The authors acknowledge the contributions of colleagues in the metadata groups of RDA and particularly: Jane Greenberg, Alex Ball, Bridget Almas, Sayeed Choudhury, David Durbin.

Links:

http://dublincore.org/metadata-basics/
http://dublincore.org/
http://ckan.org/
http://www.eurocris.org/Index.php?page=CERIFreleases&t=1
https://rd-alliance.org/groups/metadata-ig.html
https://rd-alliance.org/groups/metadata-standards-directory-working-group.html
https://rd-alliance.org/groups/data-context-ig.html
https://rd-alliance.org/groups/research-data-provenance.html

References:

[1] K.G. Jeffery, A. Asserson, N. Houssos, B. Jörg: "A 3-layer model for Metadata", in CAMP-4-DATA Workshop, proc. International Conference on Dublin Core and Metadata Applications, Lisbon September 2013.
http://dcevents.dublincore.org/IntConf/dc-2013/schedConf/presentations?searchField=&searchMatch=&search=&track=32

[2] K.G. Jeffery, N. Houssos, B. Jörg, A. Asserson: "Research Information Management: The CERIF Approach", Int. J. Metadata, Semantics and Ontologies, Vol. 9, No. 1, pp 5-14 2014.

Please contact:

Keith G. Jeffery
Keith G. Jeffery Consultants, UK
E-mail:
Keith.Jeffery@keithgjefferyconsultants.co.uk

RDA: Brokering with Metadata

by Stefano Nativi, Keith G. Jeffery and Rebecca Koskela

RDA is about interoperation for dataset re-use. Datasets exist over many nodes. Those described by metadata can be discovered; those cited by publications or datasets have navigational information. Consequentially two major forms of access requests exist: (1) download of complete datasets based on citation or (query over) metadata and (2) relevant parts of datasets instances from query across datasets.

Conventional end-to-end Mediator

Client and/or server developers need to write mediation software that given two dataset models transforms instances of A to B, or B to A. Given n data models then this requires $n*(n-1)$ (i.e. almost n^2) mediation modules, as depicted in Figure 1. The programmer has to understand each data model and manually match/map attributes, specifying the algorithm for instance conversion. Clearly, this does not scale.

Mediating with canonical representation

Increasingly mediators are written using a superset canonical representation internally and instances of A and B are converted to the canonical representation C. The programmer has to understand the dataset (already understanding the canonical representation) and manually match/map attributes, specifying the algorithm for the conversion of instances. C grows with increasing datasets causing issues of backward compatibility. This technique is applicable in a restricted research domain where datasets have similar data attributes. Although the number of conversions is reduced to n, the evolution of C and software maintenance cause significant cost and the restriction of research domain precludes multidisciplinary interoperability.

Extending client-server archetype: brokering services

To address these issues, it is possible to apply a new approach [1] based on a 3-tier architecture: Client-Broker-Server (C-B-S) introducing an independent broker, as showed by Figure 2. The mediation moves from the client and server to the broker. This provides the following benefits:

- Data providers and users do not have to be ICT experts so saving time for research;
- The broker developers are data scientists; they implement a complex canonical representation maintaining that for sustainability;
- They are domain agnostic and can implement optimised mediation across domains;
- A 3-tier mediation approach is well supported by cloud-based services.

However, this approach introduces new challenges:

- Broker governance: interoperability and sharing require trust;
- Community view of interoperability: a data scientist working on different domains is required while evolution from Client-Server interoperability to the revolutionary Bring-Your-Data (BYD) approach occurs.

Nowadays, a broker implements a mediating service based on C. The following sections describe the evolution of metadata-assisted brokering.

Metadata Assisted Simple Brokering: This places dataset descriptions (metadata) outside the conversion software. Tools assist the match/map between the metadata and the generation of software to do the instance conversions. There is no automated production software for this task [2] although some prototype systems providing part-automation exist [3]. They usually provide graph representations of dataset metadata for A and B, propose matching entities/attributes of A and B and allow the data scientist to correct the proposed matches. If the equivalent attributes are of different types (unusual) or in different units / precision, transformations are suggested or input. The instance conversion software is written, partially generated or generated depending on the sophistication of the tools based on the specification. The matching is difficult if the datasets differ in character representations/languages requiring sophisticated knowledge engineering with domain multilingual ontologies. Of course this technique resurrects the $n*(n-1)$ problem with associated scalability issues.

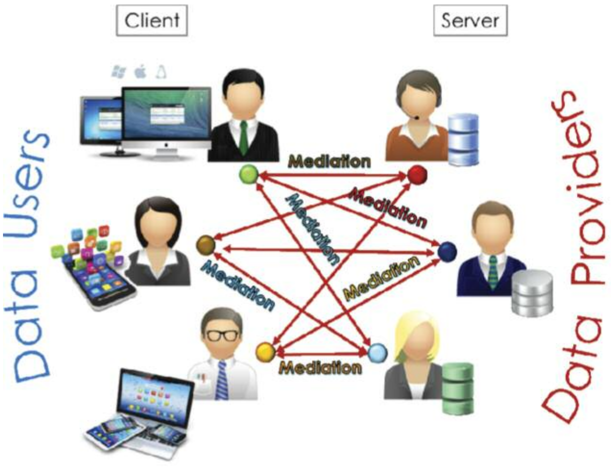


Figure 1: Conventional end-to-end mediation approach.

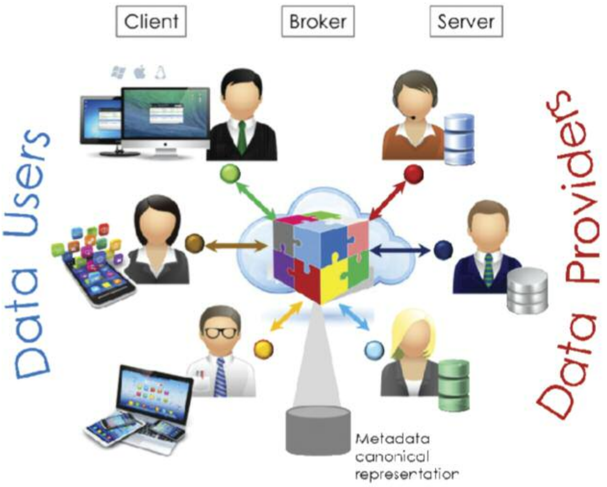


Figure 2: The C-B-S archetype.

Metadata Assisted Canonical Brokering: The next step has canonical metadata outside the conversion software with which the metadata of any dataset of interest is matched/mapped. Again, the kinds of tools mentioned above are used. This reduces the *n**(*n-1*) problem to n. However, a canonical metadata scheme for all datasets is a huge superset. Restricting the canonical metadata (based on commonality among entities/attributes) to that required for contextualisation (of which discovery is a subset) maintains the benefits of the reduction from *n**(*n-1*) to n for much of the processing and only when connecting data to software with detailed, domain-specific schemas is it necessary to write or generate specific software.

Metadata Assisted Canonical Brokering at Domain Level: The above technique for contextual metadata may be applied at detailed level restricted to a domain. This is not applicable for multidisciplinary interoperation. In any particular domain there is considerable commonality among entities/attributes e.g. in environmental sciences many datasets have geospatial and temporal coordinates and time-series data have similar attributes (e.g. temperature and salinity from ocean sensors).

The advantage of Metadata

The use of metadata and associated matching/mapping in brokering either provides a specification for the programmer writing or generates or partially generates the mediating software. Data interoperability – researched intensively since the 1970s – remains without automated processes. Partial automation, based on contextual metadata has reduced considerably the cost and time required to provide interoperability.

Metadata Assisted Brokering in RDA

Metadata is omnipresent in RDA activities; four groups specialise in metadata as mentioned in an accompanying article. These groups are now working closely with the Brokering Governance WG (https://rd-alliance.org/groups/brokering-governance.html) to promote techniques for dataset interoperation.

Link:
https://rd-alliance.org/groups/brokering-governance.html

References:

[1] S. Nativi, M. Craglia, J. Pearlman: "Earth Science Infrastructures Interoperability: The Brokering Approach", IEEE JSTARS, Vol. 6 N. 3, pp. 1118-1129, 2013.
[2] M. J. Franklin, A. Y. Halevy, D. Maier: "A first tutorial on dataspaces", PVLDB 1(2): 1516-1517 (2008). http://www.vldb.org/pvldb/1/1454217.pdf
[3] K. Skoupy, J. Kohoutkova, M. Benesovsky, K.G. Jeffery: ""Hypermedata Approach: A Way to Systems Integration", in proc. of ADBIS'99, Maribor, Slovenia, 1999, ISBN 86-435-0285-5, pp 9-15.

Please contact:

Stefano Nativi
CNR, Italy
E-mail: stefano.nativi@cnr.it

Asking the Right Questions - Query-Based Data Citation to Precisely Identify Subsets of Data

by Stefan Pröll and Andreas Rauber

Data underpins most scientific endeavours. However, the question of how to enable scalable and precise citation of arbitrary subsets of static and specifically dynamic data still constitutes a non-trivial challenge.

Although data has been the source of new knowledge for centuries, it has never received the same attention as the publications about the derived discoveries. Only recently has it been recognized as a first-class citizen in science, earning equal merit (see Link JDDCP below). Beyond mere referencing for the purpose of acknowledging the creators, it is the underlying basis and evidence for many scientific discoveries. With the increasing focus on repeatability and verifyability in the experimental sciences, providing access to the underlying data is becoming essential.

Data used to be condensed into human readable form, by aggregating source data into tables and graphs. Alternatively, the specific subset and version of data used in a study was deposited in a repository for later re-use. With the arrival of data driven science [1], the increasing amount of data processed and the increasing dynamics of data, these conventional approaches are no longer scalable.

Research datasets can be huge in terms of contained records. Scientists are often interested in a particular view of their data, using subsets tailored to a specific research question. An experiment can only be reproduced and verified if the same subset can be retrieved later. Depositing the specific subset in a repository does not scale to big data settings. Also providing the metadata helping users to find, interpet and access specific datasets again can be a challenge. Textual descriptions of subsets are hardly precise enough, require human intervention and interpretation of ambiguous descriptions in re-creating the dataset, limiting reproducibility of experiments and re-use in meta-studies.

Furthermore, many research datasets are highly dynamic: new data is added continuously via data stream, rendering conventional versioning approaches useless unless one wants to revert to old-fashioned time-delayed batch releases of annual/quarterly versions of the data. Additional dynamics arise from the need to correct errors in the data, removing erroneous data values, or re-calibrating and thus re-computing values at later points in time. Thus, researchers require a mechanism to retrieve a specific state of the data again, in order to compare the results of

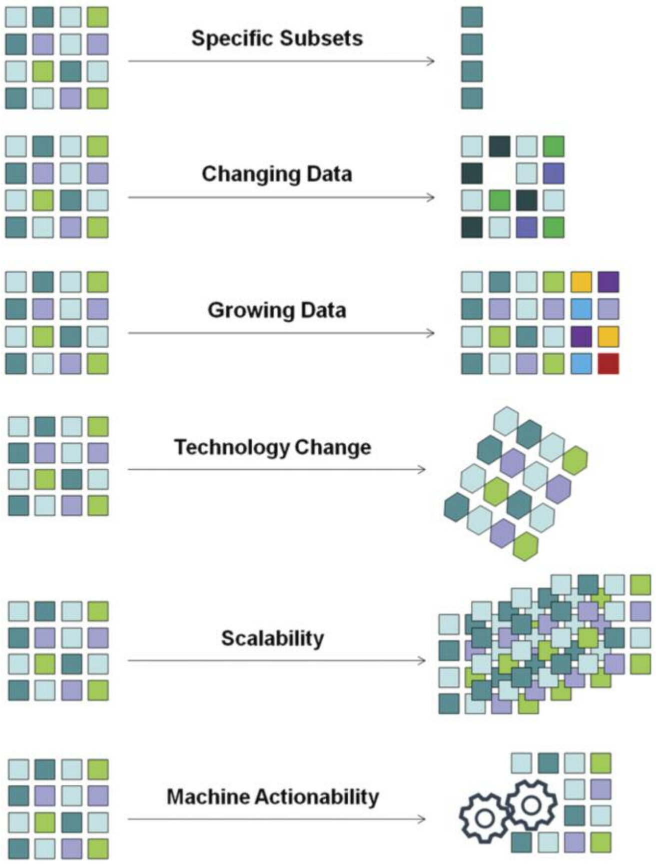


Figure 1: requirements for data citation.

previous iterations of an experiment. Obviously, storing each individual modification as a new version in a data repository does not scale for large data sources with high change frequencies.

Thus we need to devise new ways of precisely identifying specific subsets of data in potentially highly dynamic settings that do not require human intervention to interpret and assist with identifying the data as used in a specific study. The method needs to be scalable to high-volume data settings, be machine-actionable and resilient to technological changes in order to support long-term access to data. Figure 1 shows the requirements for data citation. This will support reproduction and verification of scientific experiments as well as re-use in follow-up or meta-studies, while providing citability for giving credit to the creators.

To meet these requirements we introduced a query centric view on data citation based on the tools already used by scientists during the subset creation process[2]. Our approach is based on versioned and timestamped data, where all inserts, updates and deletes of records within a dataset are recorded. Thus any state of the data source can be retrieved later. This time-stamping and versioning is often already in place in many data centres dealing with dynamic data.

To support subset identification we record the query which is used to create the subset by selecting only specific records from a data source. We trace all selection, filtering and sorting operations in an automated fashion and store these parameters together with the execution time of the query. By re-executing the time-stamped query on the time stamped and versioned data, the exact same subset can be retrieved at a later point in time. Assigning persistent identifiers to the query, each dataset can be uniquely referenced.

This approach not only allows retrieval of the original dataset, but also identifies which steps have been performed in order to create a dataset in the first place. It does implicitly provide valuable provenance information on the data set. It furthermore allows re-execution of the query against the current timestamp, re-creating an earlier dataset while incorporating all corrections made or new data added since the original study, whilst still satisfying the original selection criteria. This also supports analysis of any differences in information available between the original study and a later re-evaluation - features that cannot be easily achieved with conventional deposit-based approaches.

We first analyzed the requirements for dynamic data citation capable research data stores in the context of the APARSEN project. We implemented a reference framework in the TIMBUS Project [3], where our approach was used for describing datasets used in reports for critical infrastructure. In the SCAPE Project, we adapted an existing query interface for storing the user input and provide citation texts which could be directly used in publications. These concepts are further elaborated in the Data Citation Working Group of the Research Data Alliance (RDA). In focused workshops and pilots the approach is now being validated in diverse settings for different types of data ranging from structured SQL via XML, linked data to comma-separated value files.

Links:

JDDCP - Joint Declaration of Data Citation Principles:
https://www.force11.org/datacitation
APARSEN Project:
http://www.alliancepermanentaccess.org/index.php/aparsen/
TIMBUS Project
http://timbusproject.net/
SCAPE Project http://www.scape-project.eu/

References:

[1] T. Hey, S. Tansley, K. Tolle, eds.: “The Fourth Paradigm: Data-intensive Scientific Discovery”, Microsoft Research 2009.
[2] S. Proell, A. Rauber: “Citable by Design - A Model for Making Data in Dynamic Environments Citable”, in DATA 2013, 2013.
[3] S. Proell, A. Rauber: “Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation”, in IEEE BigData 2013, 2013.

Please contact:

Stefan Proell, SBA Research, Austria
Tel: +43150536881803
E-mail: sproell@sba-research.org

Capturing the Experimental Context via Research Objects

by Catherine Jones, Brian Matthews and Antony Wilson

Data publication and sharing are becoming accepted parts of the data ecosystem to support research, and this is becoming recognised in the field of 'facilities science'. We define facilities science as that undertaken at large-scale scientific facilities, in particular neutron and synchrotron x-ray sources, although similar characteristics can also apply to large telescopes, particle physics institutes, environmental monitoring centres and satellite observation platforms. In facilities science, a centrally managed set of specialized and high value scientific instruments is made accessible to users to run experiments which require the particular characteristics of those instruments

The institutional nature of the facilities, with the provision of support infrastructure and staff, has allowed the facilities to support their user communities by systematically providing data acquisition, management, cataloguing and access. This has been successful to date; however, as the expectations of facilities users and funders develop, this approach has its limitations in the support of validation and reuse, and thus we propose to evolve the focus of the support provided.

A research project produces many outputs during its lifespan; some are formally published, some relate to the administration of the project and some will relate to the stages in the process. Changes in culture are encouraging the re-use of existing data which means that data should be kept, discoverable and useable, for the long term. For example, a scientist wishing to reuse data may have discovered the information about the data from a journal article; but to be able to reuse this data they will also need to understand information about the analysis done to produce the data behind the publication. This activity may happen years after the original experiment has been undertaken and to achieve this, the data digital object and its context must be preserved from the start.

We propose that instead of focussing on traditional artefacts such as data or publications as the unit of dissemination, we elevate the notion of experiment or 'investigation' as an aggregation of the artefacts and supporting metadata surrounding a particular experiment on a facility to a first class object of discourse, which can be managed, published and cited in its own right. By providing this aggregate 'research object', we can provide information at the right level to support validation and reuse, by capturing the context for a given digital object and also preserving that context over the long term for effective preservation. In the SCAPE project , STFC has built on the notion of a Research Object which enables the aggregation of information about research artefacts. These are usually represented as a Linked Data graph; thus RDF is used as the underlying model and representation, with a URI used to uniquely identify artefacts, and OAI-ORE used as a aggregation container, with standard vocabularies for provenance citation and for facilities investigations. Within the SCAPE project, the focus of the research lifecycle is the experiment undertaken at the ISIS Neutron Spallation Facility. By following the lifecycle of a successful beam line application, we can collect all the artefacts and objects related to it, with their appropriate relationships. As this is strongly related to allocation of the resources of the facility, this is a highly appropriate intellectual unit for the facility; the facility want to record and evaluate the scientific results arising from the allocation of its scarce resources.

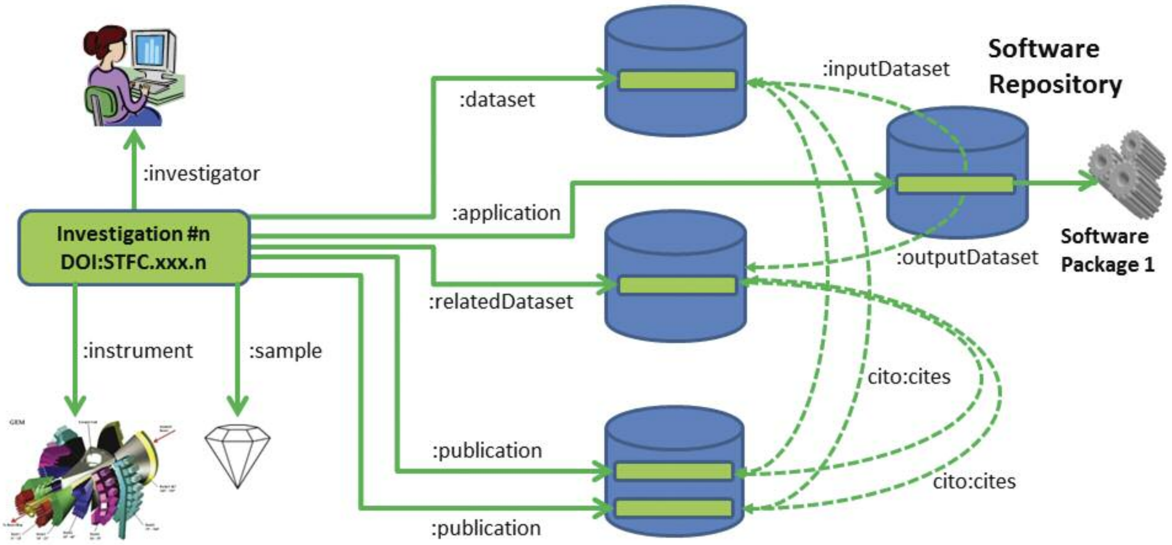


Figure 1: Schematic of an Investigation Research Object.

In this process, we provide the data in the context in which has it has been collected, and thus we are using the data provenance, in the broad sense who has undertaken the experiment and why, and how the data has subsequently been processed to provide derived data products and presentations.

Within SCAPE, STFC has explored how to construct and maintain Investigation Research Objects as part of the Facilities scientific lifecycle, using them to prototype a notion of a ‘Data Journal’, so that Experiments can be published with their full context and supporting artefacts. These can then be used as to form an archival information package for preservation of the experiment in context. Further work combines the Open Annotation Model with an ontology of experimental techniques to provide indexing of the investigations.

Future work will consider how to bring this notion into practice, particularly in support of publication data within the article preparation process, automatically combining data from different sources and relating these to the original experiment.

Links:
http://www.researchobject.org
http://www.scape-project.eu/

References:

[1] V. Bunakov, C. Jones, B. Matthews: “Towards the Interoperable Data Environment for Facilities Science”, in Collaborative Knowledge in Scientific Research Networks, ed. P. Diviacco et al., 127-153 (2015), doi:10.4018/978-1-4666-6567-5.ch007

[2] B. Matthews, V. Bunakov, C. Jones, S. Crompton: “Investigations as research objects within facilities science”, in 1st Workshop on Linking and Contextualizing Publications and Datasets, Malta, 26 Sep 2013.

Please contact:

Brian Matthews, STFC, UK
E-mail: brian.matthews@stfc.ac.uk

Engineering the Lifecycle of Data Sharing Agreements

Mirko Manea and Marinella Petrocchi

Sharing data among groups of organizations and/or individuals is essential in a modern web-based society, being at the very core of scientific and business transactions. Data sharing, however, poses several problems including trust, privacy, data misuse and/or abuse, and uncontrolled propagation of data. We describe an approach to preserve privacy whilst data sharing based on scientific Data Sharing Agreements (DSA).

The EU FP7 funded project Coco Cloud (Confidential and Compliant Clouds) is a three-year collaborative project, which started in November 2013. The project aims to facilitate data sharing in the Cloud by providing end-to-end data centric security from the client to the cloud, based on the automated definition and enforcement of Data Sharing Agreements (DSA).

Coco Cloud focuses, in part, on a case study provided by the Quiron Spanish hospital group. A common practice at Quiron is to sign agreements with external doctors, seen as ‘small hospitals’ that generate their own sensitive data whilst at the same time accessing patients’ data stored on the Quiron Cloud. The main purposes of this data sharing are: i) to provide health information to the physicians treating the hospital’s patients and ii) to refine diagnoses and therapies by including additional opinions on patients’ medical data from specialists and healthcare researchers.

Traditionally, hospitals use legal documents to regulate the terms and conditions under which they agree to share data. A key problem in the digital world is that the constraints expressed in such contracts remain inaccessible from the software infrastructure supporting the data sharing and management processes [1]. Coco Cloud approaches this issue by adopting electronic DSA.

An electronic DSA is a human-readable, yet machine-processable contract, regulating how organizations and/or individuals share data. A DSA consists of:

- Predefined legal background information, like traditional contracts. A subject matter expert (e.g., a lawyer) usually provides this information. Such data is unstructured, i.e., not organized in a predefined manner.
- Structured user-defined information, including the validity period, the involved parties, the data and, most importantly, the policy rules that constrain how data can be shared among the parties. Domain experts and end users compile these fields.

One of the key areas for Coco Cloud is the design and implementation of the DSA infrastructure, from the definition of DSA templates (predefined drafted agreements encoding, for instance, rules of law) to DSA disposal [2]. A DSA has a complex lifecycle, consisting of several stages, as depicted in Figure 1.

The template definition stage is a preliminary phase in which a pool of available DSA templates is created, according to the purpose of the data sharing and the data classification to be regulated by the DSA.

The authoring stage is an editing tool-assisted phase during which the stakeholders prepare the DSA itself. The result of the authoring stage is an electronic, human readable DSA document.

The human readable DSA is then translated into a machine-readable document with rules for verification and formal checking in the analysis stage. The authoring and analysis stages are iterative: they are repeated until the DSA rules satisfy all the required properties being checked (e.g., conflicting rules

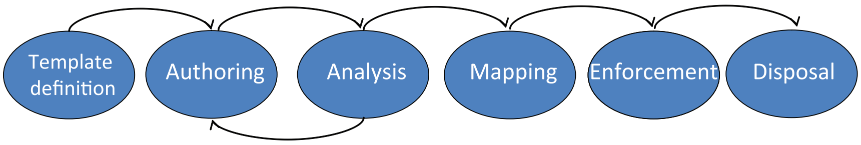
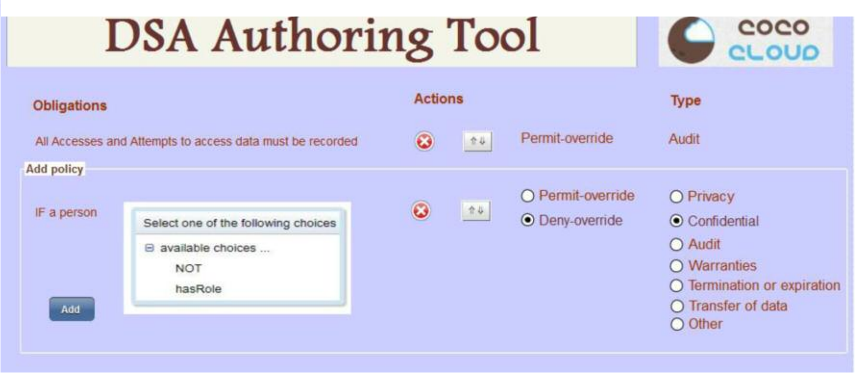


Figure 1: DSA Lifecycle.



DSA Authoring Tool

coco cloud

Obligations	Actions			Type
All Accesses and Attempts to access data must be recorded	☒	↑↓	Permit-override	Audit

Add policy

IF a person

Select one of the following choices

⊟ available choices ...

NOT

hasRole

Add

○ Permit-override
● Deny-override

○ Privacy
● Confidential
○ Audit
○ Warranties
○ Termination or expiration
○ Transfer of data
○ Other

Figure 2: DSA Authoring Tool.

checking). The DSA rules are then translated to a set of enforceable security policies during the mapping stage. The enforcement stage is the phase in which the DSA is enacted on the specific data being shared. A DSA enters the final disposal stage when the contracting parties agree that this DSA is no longer useful.

During the first year of the project, activities concentrated on: 1) design of a user-friendly authoring tool, guiding the users throughout DSA definition; 2) formalizing the agreement writing by programmatically encoding the typical sections that lawyers currently embed in paper; 3) studying the applicable Terms of Law (both national and international), to define the legal constraints that must hold when scientific or medical data are to be shared and used within a community.

In particular, we have proposed a `three-step authoring phase'. In step 1, legal experts populate a DSA template, encoding the applicable legal policies (e.g., EU Directive 95/46/EC on personal data protection). In step 2, domain experts define a context-specific policy (e.g., healthcare policy professionals define the organization-specific constraints for medical data subject to scientific investigations). Finally, in step 3, the end-users optionally fill some input forms to set their privacy preferences (e.g., consenting to the processing of their data).

The Coco Cloud project partners are HP Italy (coordinator), the Italian National Research Council, SAP, Imperial College London, Bird & Bird, ATOS, University of Oslo, AGID, and Grupo Hospitalario Quiron.
Future work will investigate ways to easily define policies for a specific domain or context (e.g., healthcare or government). We are planning to use standard Web ontologies (e.g., SNOMED CT [3] for healthcare) to define domain vocabularies and leverage them to implement an authoring tool that is easy-to-use but able to express semantically sound policy rules with guiding wizards. A first mockup of the DSA authoring tool is shown in Figure 2.

Link:
http://www.coco-cloud.eu/

References:

[1] M. Casassa-Mont et al.: “Towards safer information sharing in the Cloud”, Intl. Journal of Information Security, Springer, August 2014.
[2] R. Conti et al.: “Preserving Data Privacy in e-Health”, Engineering Secure Future Internet Services and Systems, Springer, 2014, 366-392.
[3] International Health Terminology Standards Development Organisation (IHTSDO), http://ihtsdo.org/snomed-ct/, 2014

Please contact:

Marinella Petrocchi
IIT-CNR, Italy
E-mail: marinella.petrocchi@iit.cnr.it

Mirko Manea
Hewlett-Packard Italiana, Italy
E-mail: mirko.manea@hp.com

Cross-disciplinary Data Sharing and Reuse via gCube

by Leonardo Candela and Pasquale Pagano

Data sharing has been an emerging topic since the 1980's. Science evolution – e.g. data-intensive, open science, science 2.0 – is revamping this discussion and calling for data infrastructures capable of properly managing data sharing and promoting extensive reuse. 'gCube', a software system that promotes the development of data infrastructures, boasts the distinguishing feature of providing its users with Virtual Research Environments where data sharing and reuse actually happens.

gCube - a software system designed to enable the creation and operation of an innovative typology of data infrastructure - leverages Grid, Cloud, digital library and service-orientation principles and approaches to deliver data management facilities as-a-service. One of its distinguishing features is that it can serve the needs of diverse communities of practice by providing each with one or more dedicated, flexible, ready-to-use, web-based working environments, i.e. Virtual Research Environments [1].

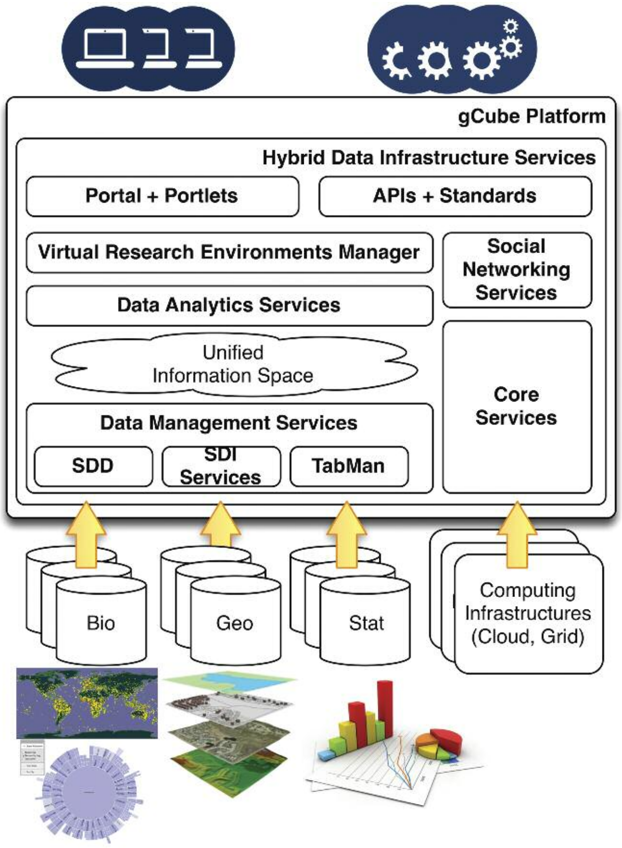

gCube Platform
Hybrid Data Infrastructure Services
Portal + Portlets
APIs + Standards
Virtual Research Environments Manager
Social Networking Services
Data Analytics Services
Unified Information Space
Core Services
Data Management Services
SDD
SDI Services
TabMan
Bio
Geo
Stat
Computing Infrastructures (Cloud, Grid)

Figure 1: The gCube System Architecture.

gCube provides its users with services for seamless access to species data, geospatial data, statistical data and semi-structured data from diverse data providers and information systems. These services can be exploited both via web-based graphical user interfaces and web-based protocols for programmatic access, e.g., OAI-PMH, CSW, SDMX.

For species data, gCube is equipped with a Species Data Discovery (SDD) Service [2] which mediates over a number of data sources including taxonomic information, checklists and occurrence data. The service is equipped with plug-ins interfacing with major information systems such as Catalogue of Life, Global Biodiversity Information Facility, Integrated Taxonomic Information System, Interim Register of Marine and Nonmarine Genera, Ocean Biogeographic Information System, World Register of Marine Species. To expand the number of information systems and data sources integrated into SDD, the VRE data manager can simply implement (or reuse) a plug-in. Each plug-in can interact with an information system or database by relying on a standard protocol, e.g., TAPIR, or by interfacing with its proprietary protocol. Plug-ins mediate queries and results from the language and model envisaged by SDD to the requirements of a particular database. SDD promotes a data discovery mechanism based on queries containing either the scientific name or common name of a species. Furthermore, to tackle issues arising from inconsistency in taxonomy among data sources, the service supports an automatic query expansion mechanism, i.e. the query could be augmented with ‘similar’ species names. Discovered data is presented in a homogenized form, e.g., in a typical Darwin Core format.

For geospatial data, gCube is equipped with services generating a Spatial Data Infrastructure compliant with OGC standards. In particular, it offers a catalogue service enabling the seamless discovery of and access to every geospatial resource registered or produced via gCube services. These resources include physical and biochemical environmental parameters, such as temperature and chlorophyll, species distribution and occurrence maps, and other interactive maps. Some of these resources are obtained by interfacing with existing Information Systems including FAO GeoNetwork, myOcean and World Ocean Atlas. New resources can be added by linking data sources to the SDI via standards or ad-hoc mediators. On top of the resulting information space, gCube offers an environment for identifying resources and overlays them through an innovative map container that caters for sorting, filtering, and data inspection further to standard facilities such as zoom in.

For statistical data, the infrastructure is equipped with a dedicated statistical environment supporting the whole lifecycle of statistical data management, including data ingestion, curation, analysis and publication. This environment provides its users with facilities for creating new datasets and code lists by using sources like CSV or an SDMX repository, curating the datasets (by using controlled vocabularies and code lists, defining data types and correcting errors), manipulating the datasets with standard operations like filtering, grouping, and aggregations, analysing the datasets with advanced mining techniques, such as trend and outlier detection, producing graphs from the datasets, and finally publishing datasets in an SDMX registry for future use.
On top of the unified information space which is underpinned by the facilities described above, gCube provides its users with social networking facilities [2] and data analytics facilities [3].

Link:
http://www.gcube-system.org

References:

[1] L. Candela, D. Castelli, P. Pagano (2013) Virtual Research Environments: an Overview and a Research Agenda. Data Science Journal, 12:GRDI75–GRDI81, 2013.
[2] M. Assante et al. (2014) A Social Networking Research Environment for Scientific Data Sharing: The D4Science Offering. The Grey Journal, Vol. 10, Number 2, 2014 □
[3] L. Candela et al. (2014) An infrastructure-oriented approach for supporting biodiversity research. Ecological Informatics, DOI: 10.1016/j.ecoinf.2014.07.006, Elsevier

Please contact:

Pasquale Pagano
ISTI-CNR, Italy
E-mail: pasquale.pagano@isti.cnr.it

Toward Automatic Data Curation for Open Data

by Thilo Stadelmann, Mark Cieliebak and Kurt Stockinger

In recent years large amounts of data have been made publicly available: literally thousands of open data sources exist, with genome data, temperature measurements, stock market prices, population and income statistics etc. However, accessing and combining data from different data sources is both non-trivial and very time consuming. These tasks typically take up to 80% of the time of data scientists. Automatic integration and curation of open data can facilitate this process.

Most open data has scientific or governmental origins and much of it resides in isolated data stores. In data warehousing, data integration as a discipline provides best practices concerning data preparation and data management tasks by offering standardized processes and tool chains. However, with the recent popularity of Big Data, an unprecedented number of new data sources contribute to an increasingly heterogeneous trove of data. Hence, 'data curation' – a fully automated means of intelligently finding and combining possible data sources in the unified space of internal and open data sources – is in high demand [1].

We recently finished a market research and architectural blueprint, funded by the Hasler Stiftung, to evaluate requirements and business cases concerning the development of such an automatic data curation system in Switzerland.

Market Research

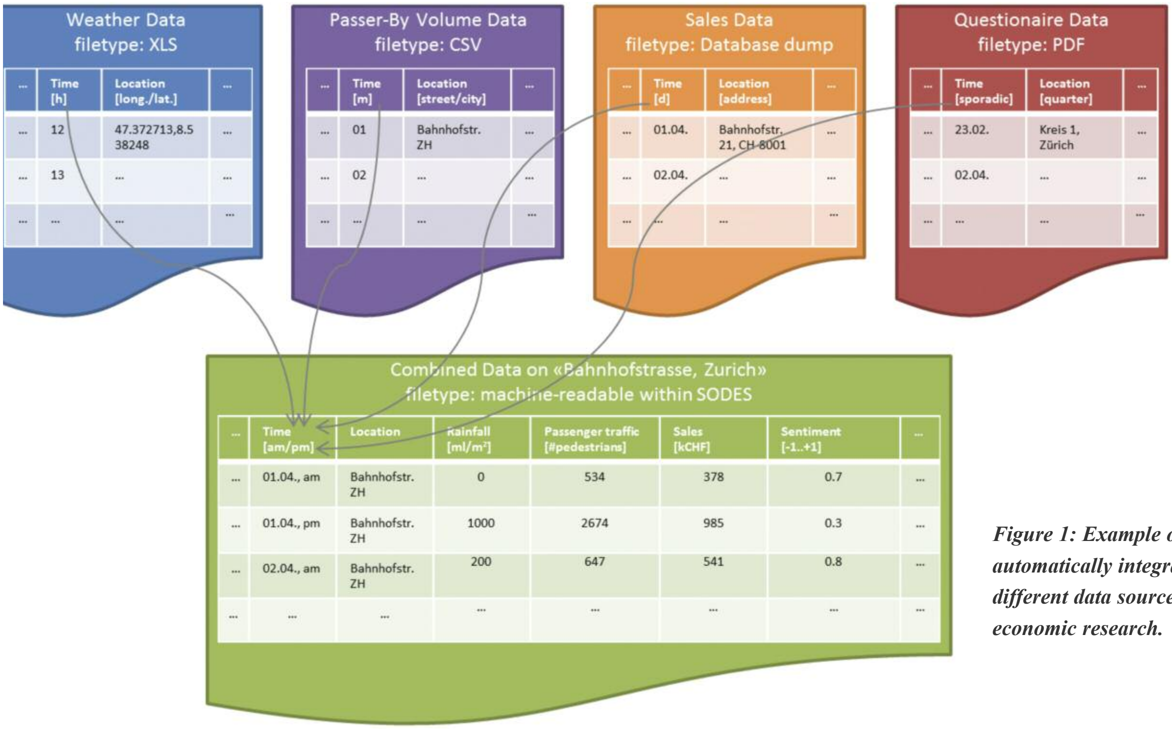
The Swiss ICT sector is healthy and innovative, sustained by strong players in research (e.g., universities and privately held research institutions) and industry (e.g., finance and pharma) as well as a large ecosystem of SMEs and startups. All surveyed parties among this group of stakeholders responded by stating their need for better data curation support of open data that has not been previously integrated within internal data sources.

As examples, we identified several use cases that rely on the existence of such a service:

- Economic research would be significantly improved by using automatic data curation for unifying tax and rent data of Swiss municipalities
- In a transportation research project, the overall project cost increased by 25% because of scattered and heterogeneous public data.
- Scientific digital media archives could fully index their work, thereby creating new research and application possibilities.

For these reasons, national funding agencies are very keen to support the development of such a service based on a solid model of business and operations.

Such a business model could consist of offering automatic data curation as software-as-a-service for open data. In order to comply with open data standards, the access to the data itself has to be free, while additional services could be offered on a freemium basis. To be of interest to industrial customers, private installations to curate confidential internal data could also be offered.



Weather Data
filetype: XLS

...	Time [h]	Location [long./lat.]	...
...	12	47.372713,8.538248	...
...	13	...	...
...	...	...	...

Passer-By Volume Data
filetype: CSV

...	Time [m]	Location [street/city]	...
...	01	Bahnhofstr. ZH	...
...	02	...	...
...	...	...	...

Sales Data
filetype: Database dump

...	Time [d]	Location [address]	...
...	01.04.	Bahnhofstr. 21, CH-8001	...
...	02.04.	...	...
...	...	...	...

Questionaire Data
filetype: PDF

...	Time [sporadic]	Location [quarter]	...
...	23.02.	Kreis 1, Zürich	...
...	02.04.	...	...
...	...	...	...

Combined Data on «Bahnhofstrasse, Zurich»
filetype: machine-readable within SODES

...	Time [am/pm]	Location	Rainfall [ml/m²]	Passenger traffic [#pedestrians]	Sales [kCHF]	Sentiment [-1..+1]	...
...	01.04., am	Bahnhofstr. ZH	0	534	378	0.7	...
...	01.04., pm	Bahnhofstr. ZH	1000	2674	985	0.3	...
...	02.04., am	Bahnhofstr. ZH	200	647	541	0.8	...
...	...	...	...	...	...	...	...

Figure 1: Example of automatically integrating four different data sources for socio-economic research.

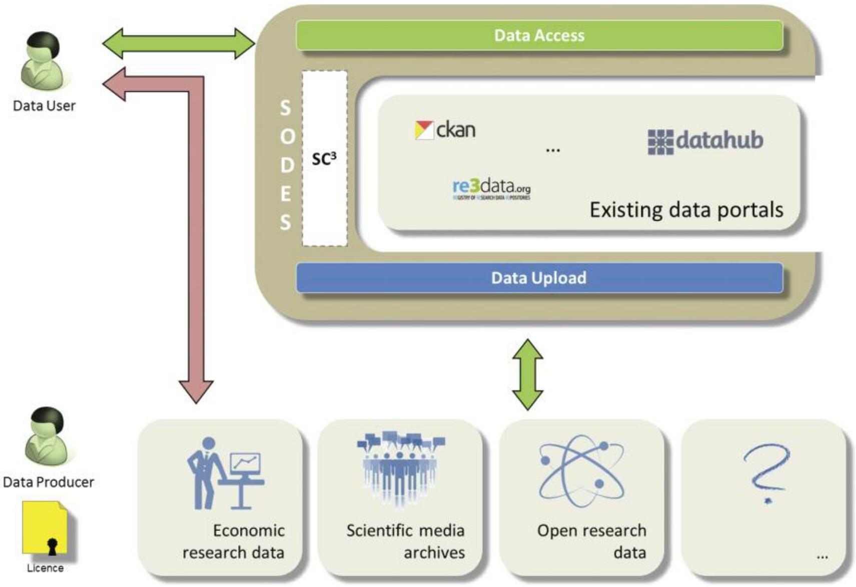


Figure 2: The high-level architecture of SODES in interaction with data providers and data portals.

While academic institutions see their general mandate to offer such services to Swiss researchers, they typically do not target industry use cases. The business models of current cloud computing providers do not necessarily coincide with offering a data curation platform to researchers, although synergies exist, e.g. through the FIWARE project.

Architecture Blueprint

We thus propose SODES – the blueprint for a Swiss Open Data Exploration System - that enables easy and intuitive access, integration and exploration of different data sources. Figure 1 shows an example of the effect: Four different data sources on Zurich's Bahnhofstrasse are automatically integrated based on common 'Time' and 'Location' columns despite the different data types and granularities therein.

SODES is designed as a platform that offers content-based search, automatic data curation and integrated data preview on top of established data technologies such as CKAN or Linked Open Data. As a service on the Internet or within any organization, SODES is envisioned to enable data scientists to do most of their data exploration in one place.

SODES can be viewed as a wrapper around existing data portals that focus on collecting resources (see Figure 2): While the user is free to still retrieve the data directly from the producer, SODES particularly adds easy-to-use data upload and data access user interfaces. These are enabled through the semantic context comprehension component SC3, powered by advanced machine learning. Therefore, SODES must process the data in machine-readable form. For advanced data analytics, users download the combined data to their tool of choice or access it remotely through standard APIs.

Future Activities

SODES is driven by a consortium of research institutions (Zurich University of Applied Sciences, University of Zurich, ETH Zurich), NGOs (Foundation opendata.ch) and industry partners (Liip AG, itopia AG) and targets Swiss data scientists in academia and industry. We are open to additional partners with interests in operating the service in order to continue development.

Links:
Swiss Open Data: http://opendata.ch
FIWARE: http://www.fi-ware.org

Reference:

[1] M. Stonebraker et al.: "Data Curation at Scale: The Data Tamer System", CIDR 2013.

Please contact:
Mark Cieliebak
ZHAW School of Engineering
Tel. +41 58 934 72 39
E-mail: ciel@zhaw.ch

An Interactive Tool for Transparent Data Preprocessing

by Olivier Parisot and Thomas Tamisier

We propose a visual tool to assist data scientists in data preprocessing. The tool interactively shows the transformation impacts and information loss, while keeping track of the applied preprocessing tasks.

Data analysis is an important topic for several domains in computer science, such as data mining, machine learning, data visualization and predictive analytics. In this context, scientists aim at inventing new techniques and algorithms to handle data and identify meaningful patterns within datasets.

Usually, data analysis techniques are worked out by using benchmark datasets: for instance, the UCI Machine Learning Repository contains a lot of material for different tasks (regressions, prediction, classification, etc.). This repository is widely used; many academic papers in the domain refer to its datasets in order to allow meaningful comparisons with the state of the art.

In practice, preprocessing is often necessary to adjust the benchmark data to the specificity of new algorithms or methods [1]. More precisely, 'data preprocessing' refers to a collection of different data transformation techniques: 'cleansing' (treatment of noise, etc.), 'dimensionality alteration' (filtering of features, etc.) and 'quantity alteration' (sampling of the data records). Moreover, a preprocessing process could drastically affect the original data, and the results of a data analysis could be substantially different depending on whether:

- outliers have been removed
- missing values have been replaced by estimations
- nominal values have been replaced by numerical values
- some columns/rows have been deleted.

A lot of uncertainty remains, related to this preprocessing step, because the modifications are not necessarily mentioned, especially in scientific publications about data analysis.

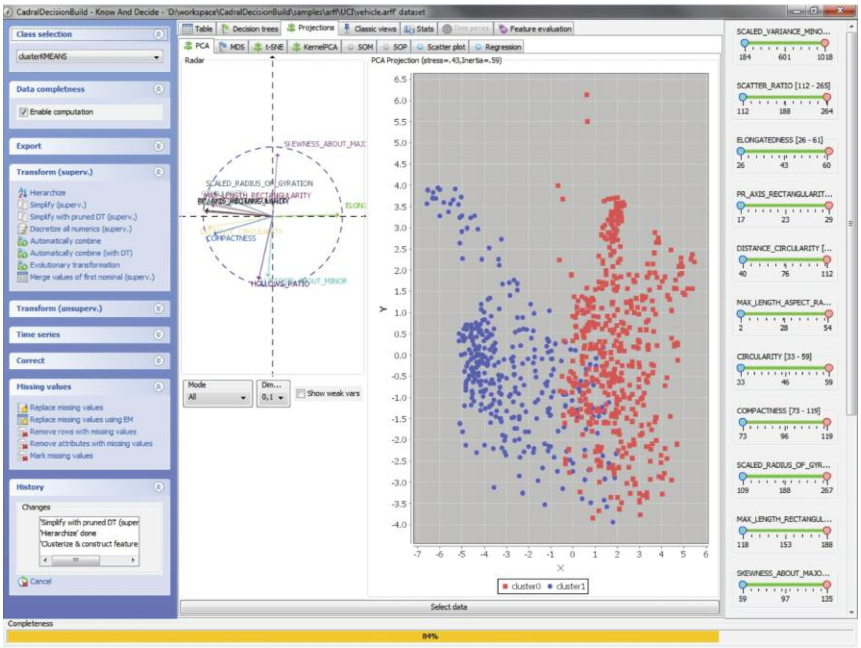
In order to improve the transparency regarding the preprocessing phase, we have developed a JAVA standalone tool that allows the data to be transformed while keeping traces of the transformations. The tool is developed on top of the WEKA library, and allows the following preprocessing operations: column/row deletion, discretization of the numerical features, feature selection, constant feature deletion, missing values imputation, outlier deletion, attribute transformation (numerical fields to nominal fields, nominal fields to binary fields, etc.).

The main features of the tool include:

- The tool facilitates interactive data transformation: in fact, the user interface provides an exploratory process to successively apply transformation operations, and then check the results using visual views, such as tables, trees, heat maps, 2D projections, etc. During this process, the user can consult the history of the applied transformations and can cancel them as required.
- After each data transformation, it is critical to gauge the intensity of data transformation (for example, the discretization of numerical values implies information loss) [2]. To this end, the tool instantly computes the ratio of values that are kept unchanged during the preprocessing steps [3]. This indicator is continuously shown in the user interface of the tool (Figure 1): 100% represents a slight transformation, 0% represents a considerable transformation.
- The tool provides support to generate a consistent sequence of data processing: for example, if a data analyst wants to normalize these data, the tool will show him that outliers and extreme values should be removed.
- In addition, the tool is able to automatically transform the data for a specific task: as an example, an algorithm has been developed in order to obtain transformed data that lead to simple prediction models [3].

The tool can be applied in several use cases to improve the scientific reusability of data. Firstly, it helps to write reusable scientific papers by providing a full description of the preprocessing steps that have been applied on this dataset; moreover, it will help to avoid the 'cherry picking issue', that biases the results of a lot of data analysis papers. Secondly, it helps data scientists to inject data assumptions into real datasets (some techniques need data without missing values, others need data with numerical values only, etc.). More precisely, it allows the transparent construction of synthetic datasets from well-known data (such as the datasets from the UCI Repository) that finally can be used in experiments. As the transformation steps and the information loss indicator are explicitly shown by the tool, they can be described in the further technical/academic papers: it will improve the reproducibility for the article's readers.

A pending issue is to deal with data for which the original source is known but the preprocessing pipeline is unknown due to a lack of documentation. As a future work, we plan to create a reengineering method in order to automatically determine the preprocessing operations that have been applied, given an original dataset and its transformed version.


CadralDecisionBuild - Know And Decide - 'D:\workspace\CadralDecisionBuild\samples\arff\UCI\vehicle.arff' dataset

Class selection: clusterKMEANS

Data completness: ☑ Enable computation

History – Changes:
'Simplify with pruned DT (super
'Hierarchize' done
'Clusterize & construct feature

Mode: All
Dim...: 0,1
□ Show weak vars

PCA Projection (stress=.43,Inertia=.59)

Completeness: 84%

SCALED_VARIANCE_MINO... 184 601 1018
SCATTER_RATIO [112 - 265] 112 188 264
ELONGATEDNESS [26 - 61] 26 43 60
PR_AXIS_RECTANGULARIT... 17 23 29
DISTANCE_CIRCULARITY [... 40 76 112
MAX_LENGTH_ASPECT_RA... 2 28 54
CIRCULARITY [33 - 59] 33 46 59
COMPACTNESS [73 - 119] 73 96 119
SCALED_RADIUS_OF_GYR... 109 188 267
MAX_LENGTH_RECTANGUL... 118 153 188
SKEWNESS_ABOUT_MAJO... 59 97 135

Figure 1: Preprocessing of the 'vehicle' dataset from the UCI repository: the data are represented using a PCA projection, the progress bar shows the data completeness and the applied operations are detailed on the bottom left of the window.

Links:
http://archive.ics.uci.edu/ml/
http://www.cs.waikato.ac.nz/ml/weka/

References:

[1] Fazel Famili et al.: "Data preprocessing and intelligent data analysis", International Journal on Intelligent Data Analysis, Volume 1, Issues 1–4, 1997.

[2] Shouhong Wang, Wang Hai: "Mining Data Quality In Completeness", ICIQ 2007, pp. 295-300.

[3] Olivier Parisot et al.: "Data Wrangling: A Decisive Step for Compact Regression Trees", CDVE 2014, pp. 60-63.

Please contact:

Olivier Parisot, Thomas Tamisier
Public Research Centre – Gabriel Lippmann, Belvaux, Luxembourg
E-mail: parisot@lippmann.lu, tamisier@lippmann.lu

e-Infrastructure across Photon and Neutron Sources

by Juan Bicarregui and Brian Matthews

Today's scientific research is conducted not just by single experiments but rather by sequences of related experiments or projects linked by a common theme that lead to a greater understanding of the structure, properties and behaviour of the physical world. This is particularly true of research carried out on large-scale facilities such as neutron and photon sources where there is a growing need for a comprehensive data infrastructure across these facilities to enhance the productivity of their science.

Photon and neutron facilities support fields as varied as physics, chemistry, biology, material sciences, energy technology, environmental science, medical technology and cultural heritage. Applications are numerous: crystallography reveals the structures of viruses and proteins important for the development of new drugs; neutron scattering identifies stresses within engineering components such as turbine blades, and tomography can image microscopic details of the structure of the brain. Industrial applications include pharmaceuticals, petrochemicals and microelectronics Research carried out at neutron and synchrotron facilities is rapidly growing in complexity. Experiments are also increasingly being carried out by international research groups and in more than one laboratory. Combined with the increased capability of modern detectors and high-throughput automation, these facilities are producing an avalanche of data that is pushing the limits of IT infrastructures.

In addition, there is a push from policy-makers and some scientific communities to make data 'open' in order to encourage transparency and sharing between scientists. It therefore makes sense to build a common data infrastructure across different photon and neutron facilities that makes data management and analysis more efficient and sustainable and maximises the science throughput of their user communities.

Established in 2008 by the ESRF, ILL, ISIS and Diamond, the PaN-data consortium now brings together 13 large European research infrastructures that each operate hundreds of instruments used by some 33,000 scientists each year. Its aim is to provide tools for scientists to interact with data and to carry out experiments jointly in several laboratories. Research undertaken by PaN-data partners show that more than 20% of all European synchrotron and neutron users make use of more than one facility (see Link below). It is therefore of considerable value to offer scientists a similar user experience at each facility, and to allow them to share and combine their data easily as they move between facilities.

At the heart of PaN-data is a federated infrastructure of catalogues of experimental data that allow scientists to perform cross-facility and cross-disciplinary research with experimental and derived data. The catalogues capture information about the experiments and associated raw data, which can then be associated with analysis processes and to the final publication of results, which feed back into new research proposals.

The first stage of the project, PaN-data Europe, which ran from 2000–2011, focussed on data policy, user information exchange, scientific data formats and the interoperation of data analysis software. Essential elements of a scientific data policy framework were agreed upon, covering aspects such as storage, access and the acknowledgement of sources.

A second project, PaN-data Open Data Infrastructure, from 2011-2014, included a common user authentication system to allow users registered at one facility to access resources across the consortium using one identity; the use of standard formats so that data generated by one instrument can be readily combined with data from others; and a federated data cataloguing system with a common metadata standard, allowing users to access data generated from different sources in a uniform way. The project also extended data management across the data continuum, into analysis processes so that users will able to trace how data are used once they have been collected. Since a data infrastructure must be sustainable, the consortium is investigating which processes and tools need to be changed to allow a facility to move towards long-term preservation, and also considering approaches to scaling data management as data acquisition rates continue to grow.

PaN-data intends to continue working together, and also extend their collaboration. The Research Data Alliance (RDA) is bringing together data managers and scientists throughout the world to work together in areas such as metadata, data publishing, digital preservation, and policy enactment. PaN-data has helped establish an RDA Interest Group on the Data Need of Photon and Neutron Science (PaNSig) to encourage the development of best practice for data management in the photon and neutron community across the world. The ultimate vision is to allow users to move within and between facilities without having to learn and use new computing systems. By allowing data to be moved, shared and mixed together across the complete life-cycle of an experiment, scientists can concentrate on getting the best science from the facility.

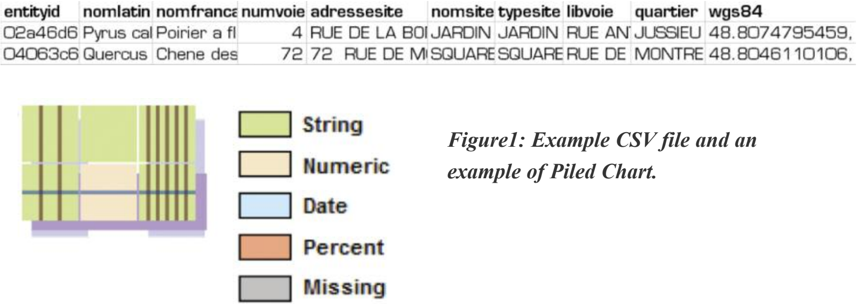
Link:
PaN-data Counting Users Survey
http://pan-data.eu/Users2012-Results

Please contact:
Brian Matthews, STFC, UK
E-mail: brian.matthews@stfc.ac.uk

Understanding Open Data CSV File Structures for Reuse

by Paulo Carvalho, Patrik Hitzelberger and Gilles Venturini

Open Data (OD) is one of the most active movements contributing to the spread of information over the web. However, there is no common standard to publish datasets. Data is made available by different kind of entities (private and public), in various formats and according to different cost models. Even if the information is accessible, it does not mean it can be reused. Before being able to use it, an aspiring user must have knowledge about its structure, location of meaningful fields and other variables. Information visualization can help the user to understand the structure of OD datasets.



entityid	nomlatin	nomfranca	numvoie	adressesite	nomsite	typesite	libvoie	quartier	wgs84
02a46d6	Pyrus cal	Poirier a fl	4	RUE DE LA BOI	JARDIN	JARDIN	RUE AN	JUSSIEU	48.8074795459,
04063c6	Quercus	Chene des	72	72 RUE DE M	SQUARE	SQUARE	RUE DE	MONTRE	48.8046110106,

Figure1: Example CSV file and an example of Piled Chart.

The Public Research Centre Gabriel Lippmann (Luxembourg), together with the University François-Rabelais of Tours (France) are currently running a project studying OD integration and how information visualization may be used to support the end user in this process. There are numerous obstacles to overcome before the great business and societal potential of OD can be realized. This project aims to address some of these obstacles.

A key problem is the plethora of available formats for OD, including: PDF, CSV, XLS, XML and others which may or may not be machine-readable. In 2011, PDF was the most common format. Since then, however, things have been changing. The use of machine-readable formats is strongly encouraged by several initiatives. The use of tabular data representation, CSV format in particular, is continually growing. Despite the increasing popularity of such files, there is a dearth of recent studies on tabular information formats and their visualization. The last relevant work in this field was published in 1994 - Table Lens visualization for tabular information [1]. Our project addresses this gap in the research, focusing on CSV format since it is currently one of the most OD used formats and the most important tabular data format.

Although CSV is a simple tabular format, its semantic and syntactic interpretation may be difficult. Since CSV does not have a formal specification, the interpretation of files can be a complex process. Nevertheless, the interpretations converge to some basic ideas described in the RFC 4180 [2]. In order to reuse OD datasets, it is mandatory to get a global idea of their structures. Information visualization may be used to create a real-time solution in order to obtain a quick and efficient global overview of OD CSV files. Our intention is to provide an effective tool to:
- estimate the size of analysed file(s);
- show the entire structure of OD CSV files;
- view what kind of data each cell contains;
- detect errors so the user can complete/correct them before use;
- detect and locate missing values;
- compare the structure of different generations of CSV files, because OD datasets are commonly published periodically.

Piled Chart has been created to achieve these objectives. Piled Chart shows the entire structure of a CSV file using a cell-based shape and a colour code. In the Piled Chart, sequences of rows with the same structure (rows with same data type in each cell) are piled (grouped) into a unique row (see Figure 1). The same principle is applied to the columns: columns with the same structure are grouped into a unique column. Furthermore, cells are colour-coded according to data type. Supported data types are currently numbers, strings, dates and percentage. A special colour is also applied for missing values. Because similar rows and columns are grouped, the CSV file can be represented using a reduced area giving the user a smaller surface to be analysed.

Piled Chart shows great promise, but still has room for improvement. For now, Piled Chart is unable to inspect several files simultaneously. This limitation is important, especially for files published periodically by the same entity (e.g., the monthly expenses of a government). We do not yet know how the system responds when large files are analysed, thus tests with large files must be performed. The problem of scalability, which applies to many algorithms in information visualisation, is challenging. The user-friendliness of the approach also needs to be evaluated; a user must be able to easily understand the information shown on the graph and the interaction must be easy and fluid. The project will continue until the end of 2016 in order to address these questions.

References:

[1] R.Rao, S. K. Card: “The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information”, in proc. of the SIGCHI conference on Human factors in computing systems (pp. 318-322), ACM, 1994.
[2] Y. Shafranovich: “Rfc 4180: Common format and mime type for comma-separated values (csv) files”, Cited on, 67, 2005.

Please contact:

Paulo Da Silva Carvalho
Centre de Recherche Public - Gabriel Lippmann
E-mail: dasilva@lippmann.lu

How Openness can Change Scientific Practice

by Robert Viseur and Nicolas Devos

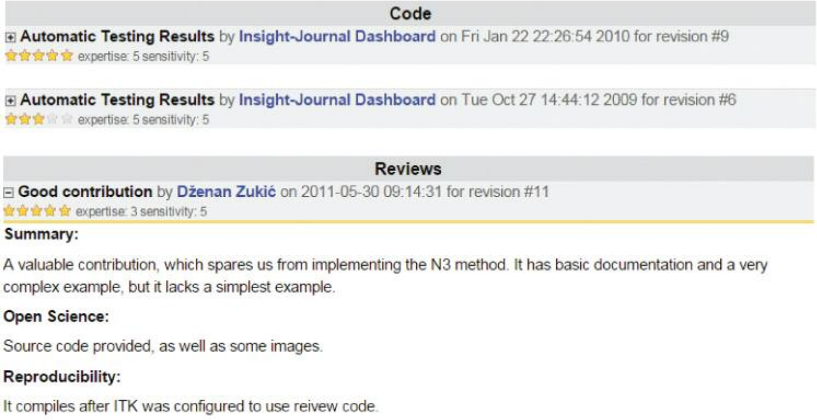
The term 'open data' refers to "information that has been made technically and legally available for reuse". Open data is currently a hot topic for a number of reasons, namely: the scientific community is moving towards reproducible research and sharing of experimental data; the enthusiasm, especially within the scientific community, for the semantic web and linked data; the publication of datasets in the public sector (e.g., geographical information); and the emergence of online communities (e.g., OpenStreetMap). The open data movement engages the public sector, as well as business and academia. The motivation for opening data, however, varies among interest groups.

The publication of scientific data is the logical continuation of open access initiatives. Such initiatives, e.g., the Budapest Open Access Initiative, wish to achieve open access to scientific publications - i.e., access without financial, legal or technical barriers. Three criteria must be met for data to be considered open: (i) access to publications should be free; (ii) the data should not be protected by digital rights management (digital handcuffs) or scrambled, (iii) access policies should be clearly communicated and allow the copy, distribution and modification of data.

Murray-Rust notes that the lack of a consistent definition of the terms 'open' and 'open access' may cause confusion. The term 'open' should imply that the scientific data belongs to the scientific community and is made freely available. The accessibility of scientific publications and the related data is of great interest both for the validation of the research and the reuse of the data in new studies that may (or may not) have been foreseen by the original authors. The practice of data sharing is more common in some scientific disciplines, such as biosciences where the data is published and consolidated into databases. Nevertheless, some publishers aggressively defend their copyright and are opposed to new scientific practices that publicly disseminate the research results with their source code, the data structures, the experimental design, the parameters, the documentation and figures [3].

The journal "Insight" shows how open access, open source and open data could change scientific practice. "Insight" is a peer reviewed online publication that is associated with the software Insight Segmentation and Registration Toolkit (ITK). ITK, which is supported by the company Kitware, is an open source software tool for image analysis. Scientific results are published with each article as per usual, but are also accompanied by the source code and the data in order to enhance the reproducibility of the research ('reproducible research') [1]. The technical infrastructure automates the source code compilation and testing. Several authors have developed the idea of 'executable papers', based on Linked Data and the Cloud infrastructure [2].

Open access to data can be fostered by governments. Governments can impose open access to research units, as illustrated by the UK Department for International Development, which in 2012, opened access to development research data in order to stimulate innovation.



Code

⊞ **Automatic Testing Results** by **Insight-Journal Dashboard** on Fri Jan 22 22:26:54 2010 for revision #9
★★★★★ expertise: 5 sensitivity: 5

⊞ **Automatic Testing Results** by **Insight-Journal Dashboard** on Tue Oct 27 14:44:12 2009 for revision #6
★★★☆☆ expertise: 5 sensitivity: 5

Reviews

⊟ **Good contribution** by **Dženan Zukić** on 2011-05-30 09:14:31 for revision #11
★★★★★ expertise: 3 sensitivity: 5

Summary:

A valuable contribution, which spares us from implementing the N3 method. It has basic documentation and a very complex example, but it lacks a simplest example.

Open Science:

Source code provided, as well as some images.

Reproducibility:

It compiles after ITK was configured to use reivew code.

Figure 1: An extract from Insight illustrating the automatic testing and contribution process.

Several issues will have to be addressed to facilitate the spread of open access to scientific data. Few peer reviewed journals support open data or possess the technical infrastructure needed to power the automation of the code execution, compilation and testing. Moreover, high ranking journals benefit from their advantageous market position and have no incentive to develop new publication methods. As a result, the mobilization of the entire scientific community, its commitment to play the openness game, and its support to high-quality innovative journals, will be fundamental to the success of these new collaborative practices.

Links:

Insight: http://www.insight-journal.org
Insight Segmentation and Registration Toolkit: http://www.itk.org
Open Knowledge Foundation: http://www.okfn.org
Panton Principles for Open scientific data: http://www.pantonprinciples.org

References:

[1] J. Jomier, A. Bailly, M Le Gall, et al.: "An open-source digital archiving system for medical and scientific research", Open Repositories, 2010, vol. 7.

[2] T. Kauppinen, G.M. de Espindola: "Linked open science-communicating, sharing and evaluating data, methods and results for executable papers", Procedia Computer Science, 2011, vol. 4, p. 726-731.

[3] V. Stoddden: "The legal framework for reproducible scientific research: Licensing and copyright", Computing in Science & Engineering, 2009, vol. 11, no 1, p. 35-40.

Please contact:

Robert Viseur, Nicolas Devos
CETIC, Belgium
E-mail: robert.viseur@cetic.be, nicolas.devos@cetic.be

European Research and Innovation

Bridging the Gap between Testing and Formal Verification in Ada Development

by Claude Marché and Johannes Kanig

Recent technological advances in formal deductive verification are benefiting industry users of programming language "Ada". Mathematical proof complements existing test activities whilst reducing costs.

The Ada programming language was the winner of a call issued by the US Department of Defence in the late 1970's, aiming at replacing the various languages that were used in mission-critical embedded software at that time. It was named in honour of Ada Lovelace, who is recognized by the scientific community as the first computer programmer. The first ANSI standard for Ada appeared in 1983.

SPARK, a subset of Ada, which has been around for nearly as long as Ada, was originally developed by Praxis (UK) and is currently co-developed by Altran UK and AdaCore (France and USA). Compared to Ada, SPARK imposes restrictions regarding the absence of name aliasing, allowing precise static analyses of data- and information-flow. These restrictions make SPARK well-suited to the development of mission-critical systems.

Unlike Ada, SPARK allows developers to attach contracts to procedures. A contract is a set of formal logic formulas expressing either requirements (pre-conditions) or guarantees (post-conditions) of the procedure to which it is attached. SPARK comes with a verification condition generator, allowing the user to statically check that contracts are fulfilled, using computer-assisted mathematical proof techniques.

SPARK has been used in several safety-critical systems, covering avionics (jet engines, air-traffic management), rail and space applications [1]. SPARK has also been used in the development of security-critical systems. Representative case studies include the NSA Tokeneer demonstrator (Microsoft Research Verified Software Milestone Award 2011) and the iFACTS system for assisting air-traffic controllers in the UK.

The AdaCore company leads the development of the GNAT Ada compiler and the many development tools around it. These tools are distributed under an open-source licence, which means no commitment for customers. Since AdaCore's business model is based on yearly subscriptions, the company must continually improve the technology and offer professional high-quality support. To maximize the potential dissemination of the innovative solutions, AdaCore also provides a free version of the tools suitable for development under the GNU public licence.

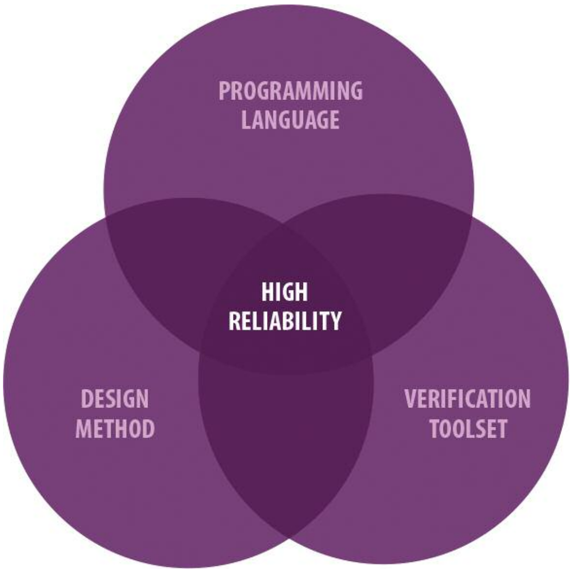


Figure 1: The SPARK2014 triptych.

Ada2012, the current version of Ada language reference, now incorporates in its core the concept of contracts in the form of regular Ada Boolean expressions rather than formal logic formulas. The primary use of such contracts is for checking their validity during execution, so that their violation can be detected by testing instead of mathematical proof.

SPARK2014 is the most recent generation of the SPARK tool suite (Figure 1). It is a full reimplementation of SPARK, developed jointly by Altran UK, AdaCore and the Toccata research team at the Inria-Saclay research centre in Orsay, France. Its main goal is to close the gap between testing and proving contracts: contracts can be checked for validity using computer-assisted proof, and testing can complement proofs when they fail [2] (Figure 2). An additional novelty is the incorporation of state-of-the-art technologies for automated theorem proving, thanks to the use of Toccata's Why3 environment and its ability to call the best SMT solvers in the market. This technology has enabled fully automatic discharge of a far greater percentage of verification conditions.

SPARK2014 will continue to evolve in the near future. A new 3-year project "ProofInUse", funded by the French National Research Agency, started in April 2014. Its aim is to promote the replacement of testing activities by proof methods to the Ada industry community. To this end, a significant effort will be made to even further automate the approach. We also plan to enlarge the subset of Ada covered by the SPARK language, enabling more applications to be developed using SPARK2014's powerful hybrid approach of test and proof. Among the new features, the SPARK verification condition generator will fully support the IEEE-754 standard for floating-point arithmetic, thanks to the expertise of the Toccata team in this domain [3]. Thus, applications making important use of numerical computations, such as in avionics and space control software, will be able to reach a very high level of assurance in the precision of computations.

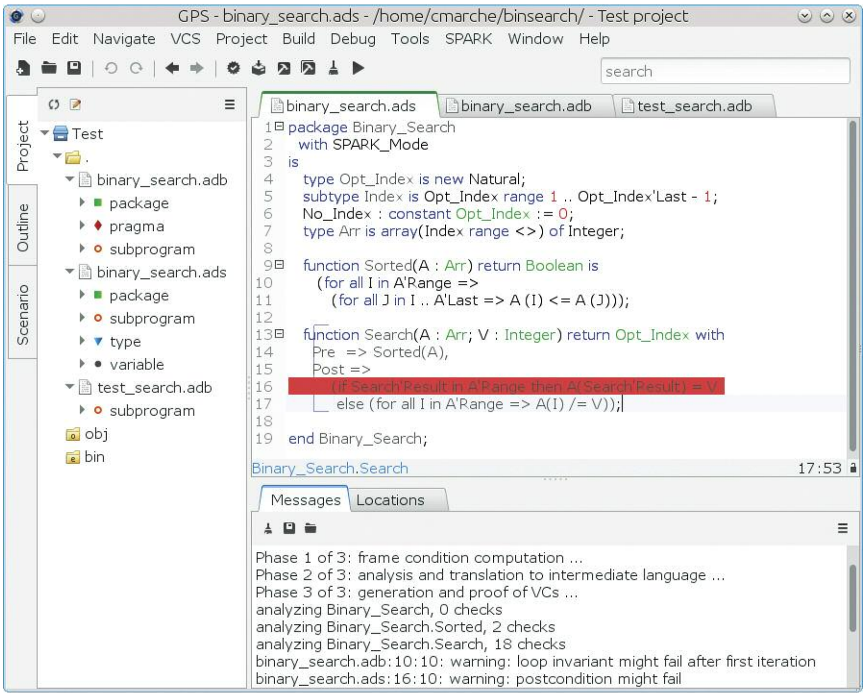
Links:
http://libre.adacore.com/
http://www.spark-2014.org/

References:

[1] R. Chapman, F. Schanda: "Are We There Yet? 20 Years of Industrial Theorem Proving with SPARK", 5th Int. Conf. on Interactive Theorem Proving, 2014.
[2] C. Dross et al.: "Rail, Space, Security: Three Case Studies for SPARK2014", 7th European Congress on Embedded Real Time Software and Systems, 2014.
[3] S. Boldo, C. Marché: "Formal verification of numerical programs: from C annotated programs to mechanical proofs", Mathematics in Computer Science, 2011.

Please contact:

Claude Marché, Inria, France
Tel: +33 172 925 969
E-mail: Claude.Marche@inria.fr



```
1  package Binary_Search
2    with SPARK_Mode
3  is
4     type Opt_Index is new Natural;
5     subtype Index is Opt_Index range 1 .. Opt_Index'Last - 1;
6     No_Index : constant Opt_Index := 0;
7     type Arr is array(Index range <>) of Integer;
8
9     function Sorted(A : Arr) return Boolean is
10       (for all I in A'Range =>
11          (for all J in I .. A'Last => A (I) <= A (J)));
12
13    function Search(A : Arr; V : Integer) return Opt_Index with
14      Pre  => Sorted(A),
15      Post =>
16          (if Search'Result in A'Range then A(Search'Result) = V
17           else (for all I in A'Range => A(I) /= V));
18
19 end Binary_Search;
```

```
Phase 1 of 3: frame condition computation ...
Phase 2 of 3: analysis and translation to intermediate language ...
Phase 3 of 3: generation and proof of VCs ...
analyzing Binary_Search, 0 checks
analyzing Binary_Search.Sorted, 2 checks
analyzing Binary_Search.Search, 18 checks
binary_search.adb:10:10: warning: loop invariant might fail after first iteration
binary_search.ads:16:10: warning: postcondition might fail
```

Figure 2: Snapshot of the GPS user interface for Ada and SPARK, displaying a program under development. The red line corresponds to a warning issued by the proof tool.

Simulations Show how Lightning Creates Antimatter

by Christoph Köhn and Ute Ebert

Active thunderstorms can emit gamma-rays and even antimatter. In fact, growing lightning channels can act as enormous particle accelerators above our heads. Researchers of the Multiscale Dynamics group at CWI have modelled and simulated these multiscale processes. They corrected previous calculations of gamma-ray emissions and, for the first time, computed the emission and propagation of positrons and neutrons. The neutrons are produced in nuclear reactions in air and might reach the ground.

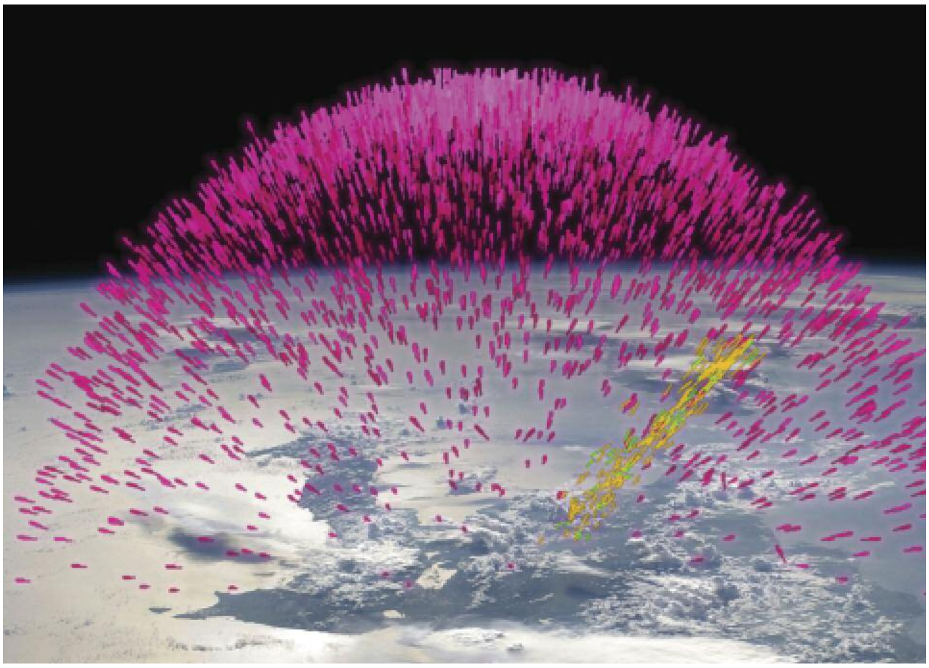
Since 1994 we have known that active thunderstorms can generate 'terrestrial gamma-ray flashes', and in 2009 NASA's Fermi satellite detected that thunderstorms can even launch beams of antimatter, namely positrons, the anti-particles of electrons (see link below). The positrons were so numerous that they could be detected at 500 km altitude by the satellite. Other researchers claim that growing lightning channels would also emit neutrons, though these observations are still under debate. Gamma rays, positrons and neutrons are typically generated in nuclear reactions – so what is going on in the clouds above our heads and around our airplanes? Should we worry? And what radiation is liberated when a lightning channel approaches ground?

The emission of particles with such high energies was not expected by lightning experts. Researchers from the Multiscale Dynamics research group at CWI, headed by Ute Ebert, were well positioned to simulate and understand this phenomenon as they were already investigating electron acceleration in technological discharges at atmospheric pressure, in the context of energy efficient plasma processing and plasma medicine and in the context of high voltage switches for electricity nets.

In essence, growing discharges in nature and technology can accelerate electrons to very high energies within an ionization wave. This process occurs very far from equilibrium in a small region around the tip of the growing channel, and the available electric energy is converted there into electron acceleration, and ionization and break-up of a small fraction of molecules, while the gas overall remains cold. If the electron energy is high enough, the electrons can generate gamma-radiation when colliding with molecules; and the gamma-radiation can subsequently create electron positron pairs and liberate neutrons and protons from the nuclei of air molecules. But the voltage in thunderstorms can reach the order of 100 MV, while the highest voltages anticipated for future long distance electricity nets are 1.2 MV. Therefore, technology oriented models have to be extended to deal with the extreme conditions of thunderstorms.

To accurately model these processes, models have to cover multiple scales in space and in energy. The CWI models start out from parameterizations of the processes that occur when electrons or other particles collide with air molecules. Some of these parameterizations had to be revised. They enter into the discharge model on the smallest scale which is of MC/PIC type, i.e., a Monte Carlo model with Particle in Cell approximation for the coupling of electric charges to the electrostatic Poisson equation [1]. In this model, the electrons follow their classical or relativistic path between air molecules, and their collisions with a random background of molecules are modelled with a Monte Carlo process. Proper averaging over the random electron motion (by taking moments of the Boltzmann equation and by truncating after the fourth moment) has delivered a set of four coupled partial differential equations for electron density, electron flux, electron energy, and electron energy flux [2]. An important and difficult aspect of both the MC/PIC model and the PDE model is the coupling of the electric charge density to the electrostatic Poisson equation. Solving the Poisson equation is a classical numerical problem. In combination with local grid refinement it is also a challenging problem, since in each time step Poisson equations are to be solved on computational grids that move in time.

A particular challenge in the present project was to also bridge the many orders of magnitude of particle energies: the


NASA's illustration how gamma-rays (pink), and electrons and positrons (yellow) are launched into space from a thunderstorm. The electrons and positrons follow the geomagnetic field lines. Credit: NASA/Goddard Space Flight Center/J.Dwyer, Florida Inst. of Technology

electric potential differences in thunderclouds are so high, that electrons can be accelerated to 40 MeV and more, which is over 100, 000, 000 times their thermal energy at room temperature (0.025 eV). This means that cross section databases for particle collisions have to span more than eight orders of magnitude, and that databases from gas discharge physics (up to 1 keV) have to be joined with those from nuclear and high energy physics (typically above 1 MeV), and the energy gap between these databases had to be filled.

Based on careful consideration of the specific processes of air molecules, and collaborating with experts from different disciplines, Christoph Köhn chose appropriate approximations within his PhD thesis [3]. This allowed him to approximate the electron acceleration ahead of a growing lightning channel, and the subsequent production of gamma-rays. Köhn was then the first to calculate the production and propagation of positrons and neutrons. Both positrons and neutrons can propagate from the thundercloud upward to the satellite, and they also can reach ground. So in terms of ground measurements it might appear as if a thunderstorm were a small nuclear reactor.

The theoretical studies were performed by Christoph Köhn within his PhD thesis under the supervision of Ute Ebert, the head of the Multiscale Dynamics research group at CWI. As a counterpart, Pavlo Kochkin investigated the emission of hard X-rays from meter long sparks in the high voltage laboratory of Eindhoven University of Technology under the supervision of Lex van Deursen. The research of both PhD students was funded by the Dutch technology foundation STW.

Links:
NASA's Fermi Catches Thunderstorms Hurling Antimatter into Space:
http://www.nasa.gov/mission_pages/GLAST/news/fermi-thunderstorms.html
Multiscale Dynamics research group at CWI:
https://www.cwi.nl/research-groups/Multiscale-Dynamics
Christoph Köhn's phd thesis:
http://www.tue.nl/en/publication/ep/p/d/ep-uid/412030/
Pavlo Kochkin's phd thesis:
http://www.tue.nl/en/publication/ep/p/d/ep-uid/413844/

References:

[1] J. Teunissen and U. Ebert: "Controlling the weights of simulation particles: adaptive particle management using k-d trees", J. Comput. Phys. 259, 318 (2014).
[2] S. Dujko, A.H. Markosyan, R.D. White and U. Ebert: "High order fluid model for streamer discharges: I. Derivation of model and transport data", J. Phys. D: Appl. Phys. 46, 475202 (2013).
[3] C. Köhn: "High-energy phenomena in laboratory and thunderstorm discharges", PhD thesis, TU Eindhoven, 2014.
http://homepages.cwi.nl/~koehn/phd_thesis/index.html

Please contact:

Ute Ebert, Christoph Köhn, CWI, The Netherlands
Tel: +31 20 592 4206/4094
E-mail: Ute.Ebert@cwi.nl, koehn@cwi.nl

GROBID - Information Extraction from Scientific Publications

by Patrice Lopez and Laurent Romary

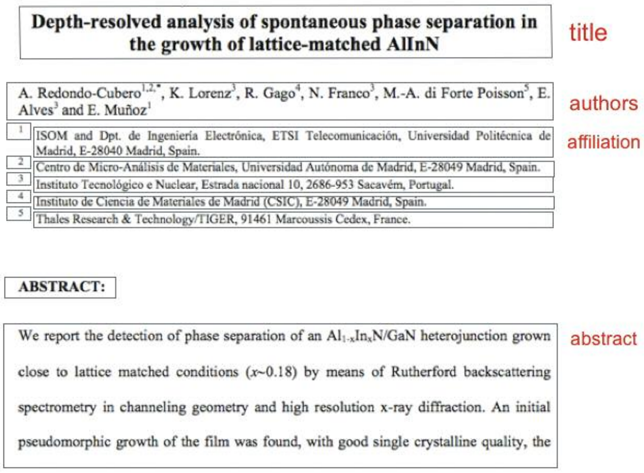
Scientific papers potentially offer a wealth of information that allows one to put the corresponding work in context and offer a wide range of services to researchers. GROBID is a high performing software environment to extract such information as metadata, bibliographic references or entities in scientific texts.

Most modern digital library techniques rely on the availability of high quality textual documents. In practice, however, the majority of full text collections are in raw PDF or in incomplete and inconsistent semi-structured XML. To address this fundamental issue, the development of the Java library GROBID started in 2008 [1]. The tool exploits "Conditional Random Fields" (CRF), a machine-learning technique for extracting and restructuring content automatically from raw and heterogeneous sources into uniform standard TEI (Text Encoding Initiative) documents.

In the worst - but common - case, the input is a PDF document. GROBID integrates fast PDF processing techniques to extract and reorganise not only the content but also the layout and text styling information. These pieces of information are used as additional features to further improve the recognition of text structures beyond the exploitation of text only information. The tool includes a variety of CRF models specialized in different sub-structures - from high level document zoning to models for parsing dates or person names. These models can be cascaded to cover a complete document.

The first and most advanced model is dedicated to the header of a scientific or technical article and is able to reliably extract different metadata information such as titles, authors, affiliations, address, abstract, keywords, etc. This information is necessary in order to identify the document, make it citable, and use it in library systems. Following an evaluation carried out for this task in 2013 by [2], GROBID provided the best results over seven existing systems, with several metadata recognized with over 90% precision and recall. For header extraction and analysis, the tool is currently deployed in the production environments of various organizations and companies, such as the EPO, ResearchGate, Mendeley and finally as a pre-processor for the French national publication repository HAL.

GROBID also includes a state of the art model for the extraction and the recognition of bibliographic citations. The references present in an article or patent are identified, parsed, normalized and can be matched with a standard reference database such as CrossRef or DocDB (patents). Citation information is considered very useful for improving search ranking and makes it possible to run bibliographic studies and graph-based social analyses. For instance, the citation notification service of ResearchGate uses GROBID bibliographic reference extraction to process every uploaded



Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

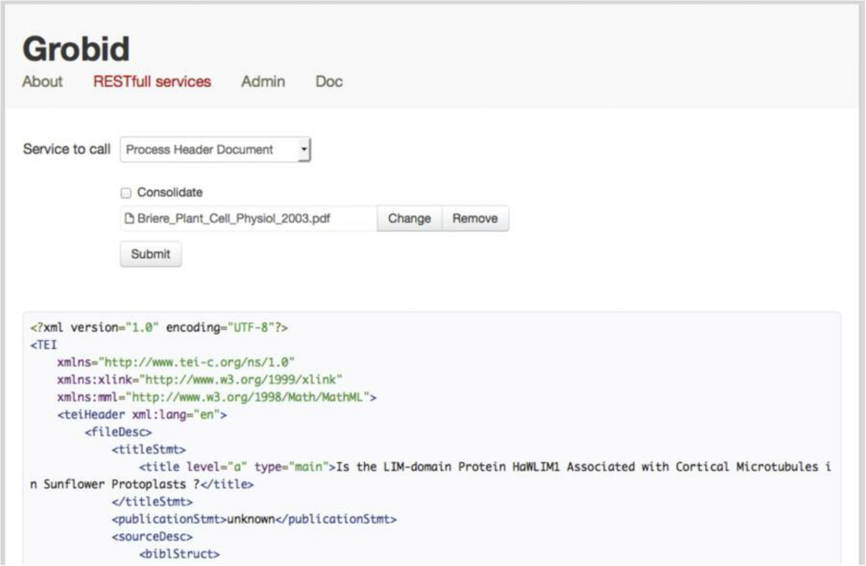
A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

1 ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.
2 Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
3 Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.
4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.
5 Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

ABSTRACT:

We report the detection of phase separation of an $Al_{1-x}In_xN$/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

Figure 1: Block segmentation of PDF documents before construing content.



Service to call: Process Header Document

□ Consolidate

Briere_Plant_Cell_Physiol_2003.pdf

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI
    xmlns="http://www.tei-c.org/ns/1.0"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:mml="http://www.w3.org/1998/Math/MathML">
    <teiHeader xml:lang="en">
        <fileDesc>
            <titleStmt>
                <title level="a" type="main">Is the LIM-domain Protein HaWLIM1 Associated with Cortical Microtubules i
n Sunflower Protoplasts ?</title>
            </titleStmt>
            <publicationStmt>unknown</publicationStmt>
            <sourceDesc>
                <biblStruct>
```

Figure 2: Online service for GROBID with TEI compliant export.

article. When an existing publication of a registered member is identified, the member can be informed where and how his work has been cited.

More challenging, the restructuring of the body of a document (potentially including figures, formula, tables, footnotes, etc.) is continually improving and is currently the object of the semi-automatic generation of more training data. Although more experimental, it can provide to a search engine for scientific literature richer and better text content and structures than basic PDF extractors (e.g., pdftotext, Apache TIKA or PDFBox).

The objectives of GROBID are still mainly research challenges, but significant efforts have also been dedicated to engineering. The tool can be used as web services or batch and is fast enough to scale to millions of documents in reasonable time and cluster. On a single low end hardware, GROBID processes, on average, three PDF documents per second or 3000 references in less than 10 seconds. Since 2011, the tool has been available as Open Source (Apache 2 licence) to any developers/third parties (see link below). New contributors are of course welcome. Version 0.3 of the tool has just been released, and its development will continue over the next few years with the participation of various national and international collaborators.

Links:
Text Encoding Initiative: http://www.tei-c.org
https://github.com/kermitt2/grobid

References:

[1] P. Lopez: "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications", in proc. of ECDL 2009, 13th European Conference on Digital Library, Corfu, Greece, 2009.
[2] M. Lipinski, et al.: "Evaluation of header metadata extraction approaches and tools for scientific PDF documents", in proc. of the 13th ACM/IEEE-CS joint conference on Digital libraries (JCDL '13), ACM, New York, NY, USA, 385-386, 2013. DOI=10.1145/2467696.2467753, http://doi.acm.org/10.1145/2467696.2467753.
[3] P. Lopez, L. Romary: "HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID", SemEval 2010 Workshop, Uppsala, Sweden. https://hal.inria.fr/inria-00493437

Please contact:

Laurent Romary, Inria, France
E-mail: laurent.romary@inria.fr

European Cyber-Security Research and Innovation

by Federico Maggi, Stefano Zanero, and Evangelos Markatos

Looking back at the evolution of cyber criminal activities, from the nineties to the present day, we observe interesting trends coming together in what may seem a perfectly orchestrated scene. In parallel with the 'security by design', we recall the importance of reactive security in a field of ever-changing arms races.

From the Morris Worm to Invisible Malware

In 1988 the Morris Worm [1] marked the beginning of the first of three decades of malicious software: malware written by developers to demonstrate their skill. In the early days, it was not uncommon to find reconnaissance traces identifying the author purposely buried in the code.

Around the beginning of the 21st century, something changed. Criminals started to see business opportunities from compromising and remotely controlling machines. Since then, opportunistic, organized and profit-driven attacks have been rising at an extraordinary pace. For the last 10–15 years the cyber criminals' goal has been to infect as many targets as possible in order to create new botnets or increase the power of those already in existence. More powerful botnets meant more profit, which came from stolen information (e.g., credentials, credit cards) or directly from renting out these attack-as-a-service infrastructures. Our analysis in Chapter 11 of the Red Book [2] shows that modern botnets are also extremely resilient, guaranteeing the cyber criminals long lasting supplies of offensive capabilities.

Today, thanks to the increased sophistication of the research and industry countermeasures, we observe a slight reduction of mass-malware attacks, which have become, to some extent, the background noise of the Internet. Meanwhile, new and more powerful actors have appeared on the scene. On the one hand, the criminal organizations are now more powerful than in the past, thanks to the technical and financial resources accumulated over the years. According to our analysis in Chapter 1, the global market of cyber crime has surpassed one trillion US dollars [3], which makes it bigger than the black market of cocaine, heroine and marijuana combined. On the other hand, hacktivists and state-sponsored attackers have skills and access to resources like never before. Our sources estimated that, as of 2012, about 88% of the EU citizens have been directly or indirectly affected by cyber-criminal activities. However, as we analyze thoroughly in Chapter 6, the era of opportunistic attacks seems to be fading, leaving the floor to high-profile persons, critical infrastructures, political activism and strategic espionage, which are now the top priority of both attackers and defenders. Modern malware samples evade automated analysis environments used in industry and research, performing only benign activities up front, stealthily interspersing unnoticeable malicious actions with benign ones.

From Incident Avoidance to Incident Response

The presence of sophisticated threats combined with this tendency to disclose vulnerabilities and an increasing value of the targeted assets obviously leads to higher levels of risk. We foresee two strategies to change this scenario and minimize the risks. The first—and perhaps not very innovative—reaction is to focus on creating less vulnerable systems by, investing in software quality, using safe programming languages, etc., and to address the remaining security bugs by creating tools and methods to find vulnerability and patch systems faster. However, experiences of recent decades have taught us that, despite significant advances in software protection, awareness among vendors, and attack-mitigation techniques, vulnerabilities are continuously being discovered. This is one of the conclusions that we draw in Chapter 4 of the Red Book, which focuses exclusively on software vulnerabilities.

What is the answer? Can we be effective in ensuring our systems' security? Our answer is that innovation in this field needs to adopt a different definition of security. A secure system today is not a perfect system, against which any attack attempt is detected and stopped before damage occurs. Vulnerabilities, attacks and incidents simply cannot be avoided. The skills, motivation, resources and persistence of modern cyber criminals are such that they will get where they want. We need to change the way we deal with the problem.

Current and Future Approaches

Incident response is not a new process, product or service. It is important to note that incident response is perhaps the most human-intensive task in system security after vulnerability research. Modern incident response should go beyond old-school control rooms with thousands of alerts and graphs calling the attention of the overwhelmed analyst. Modern incident response requires (1) extreme adaptability to new tools (e.g., malware), techniques and tactics, which change rapidly, (2) fast access to intelligence data, and (3) deep understanding of the threat scenario. Gone are the days of large, complex all-in-one security dashboards, which become immediately obsolete as the cyber criminals learn to adapt.

To complement the detailed system security research roadmap given in the Red Book, we conclude by recalling the importance of effective incident response as one of the drivers that will foster the next decade of industry and research innovation.

Link:
The SysSec Consortium: http://www.syssec-project.eu/

References:

[1] E. H. Spafford: "The Internet Worm Program: An Analysis", Purdue Technical Report CSD-TR-823, 1988, http://spaf.cerias.purdue.edu/tech-reps/823.pdf

[2] The SysSec Consortium: "The Red Book. Roadmap for Systems Security Research", http://www.red-book.eu/

[3] N. Kroes: "Internet security: everyone's responsibility", Feb. 2012, http://europa.eu/rapid/press-release_SPEECH-12-68_en.htm.

Please contact:

Federico Maggi, Politecnoco di Milano, Italy
E-mail federico.maggi@polimi.it

A Single Password for Everything?

by Jan Camenisch, Anja Lehmann, Anna Lysyanskaya and Gregory Neven

The authors have developed a three-pronged approach that can secure all of your passwords for social media, email, cloud files, shopping and financial websites, with one practically hack-proof password. This password is secured by the new "Memento protocol."

In the 2000 film "Memento" by Christopher Nolan, the protagonist suffers from short-term memory loss. Throughout the film, he meets people who claim to be his friends but, due to his condition, he never really knows whether they are truly his friends, or whether they are just trying to manipulate him or steal something from him.

This scenario got the authors thinking, because it leads to an interesting cryptographic problem: If all you can remember is a single password, then how can you store your secrets among your friends, and later recover your secrets from your friends, even if you may not remember exactly who your friends were? Or, put differently, can a user protect all her sensitive data on a set of servers with a single password, in such a way that even malicious servers do not learn anything about the data or the password when the user tries to retrieve it?

These basic questions have many applications, including protecting and recovering data on mobile devices if they are lost, encrypted data storage in the cloud, and securing access to third-party websites such as social networks, online shops, healthcare portals, or e-banking. Users nowadays are expected to remember dozens of strong, different passwords at home and in the workplace. This is obviously unreasonable, so we need a better solution.

Something important to realize about password security is that, whenever a single server can tell you whether your password is correct, then that server must be storing some information that can be used by an attacker to mount an offline dictionary attack, where the attacker simply tries to guess the password by brute force. These attacks have become so efficient lately that, if this piece of information is stolen from the server, the password itself must be considered stolen too.

The Memento protocol [1] overcomes this limitation by storing the password and data in a distributed way across multiple servers. No single server can autonomously verify a user's password; it always requires the collaboration of the other servers. To gain access, an attacker would either have to hack more than a given threshold of the servers simultaneously, or try to mount an online guessing attack on the password. The former can be addressed by using servers in different security domains and running different operating systems. The latter is prevented by letting honest servers throttle password attempts, e.g., by blocking the account after too many failed attempts, much like is done for ATM cards.

Furthermore, the Memento protocol keeps your password safe even if the user is tricked into entering her password and authenticating with a set of corrupt servers. For example, suppose you created your account on three different servers that you trust are unlikely to collude against you or to get hacked all at the same time, for example ibm.com, admin.ch, and icann.org. Next, you may be tricked in a phishing attack and you mistakenly log into ibn.com, admim.ch and ican.org. Game over for your password, right?

Wrong. With the Memento protocol, even in this situation the servers cannot figure out your password or impersonate you, because the protocol doesn't let the servers reconstruct the password when testing whether it's correct.

Instead, the protocol roughly proceeds as follows. When creating the account, the user's password p is encrypted under a special key so that at least a threshold of the servers have to collaborate to decrypt it. When logging in with password attempt q, the servers send the encryption of p back to the user, who then uses special homomorphic properties of the encryption algorithm to transform the encryption of p into an encryption of "one" if p=q, or into a an encryption or a random string if p≠q. The servers jointly decrypt the resulting ciphertext to discover whether the password was correct.

Some more cryptographic machinery is added to the protocol to obtain strong security guarantees, e.g., for the case that the user makes a typo when entering her password or that the attacker has some side information about the password, but this is the basic idea.

When using the Memento protocol, the user only needs one username and password to retrieve all her secrets. At the same time, she can rest assured that even if some of her servers get hacked or she tries to log into the wrong servers, here password and secrets remain secure.

If only the lead character in the film "Memento" had it so easy!

Link:
http://www.zurich.ibm.com/csc/security/

Reference:

[1] J. Camenisch, A. Lehmann, A. Lysyanskaya, and G. Neven, "Memento: How to Reconstruct your Secrets from a Single Password in a Hostile Environment", Advances in Cryptology – CRYPTO 2014, Springer LNCS, Volume 8617, 2014, pp 256-275.

Please contact:

Jan Camenisch, IBM Research Lab Zurich
E-mail jca@zurich.ibm.com

SIREN – A Network Infrastructure for Emergencies

by Ioannis Askoxylakis, Paschalis Papagrigoriou, Diomedes Kastanis, Panos Karampelas, and George Spanoudakis

The SIREN project (Secure, Interoperable, UAV-assisted, Rapid Emergency Deployment Communication and sensing Infrastructure) implements a secure, distributed, open, self-configured and emergency-aware network and service platform for automated, secure and dependable support of multiple mission critical applications in highly demanding and dynamic emergency environments.

In emergencies or disasters a key support factor for situation awareness, decision making and response is to provide a secure and dependable network infrastructure that aggregates connectivity over all available heterogeneous wireless broadband access technologies, including those employed for commercial network access. The infrastructure must be able to adapt to mission critical application requirements and to enable immediate and robust communications among command centres, rescue workers and the affected population.

Existing civil protection and emergency response systems do not take advantage of commercial network infrastructures, since they are not typically designed to be relied upon in the event of major extrinsic failures and are thus deemed potentially unfit to sustain the requirement of massive mission critical operations [1].

SIREN federates two existing platforms, REDComm and SecoCard, and develops a secure and dependable overlay network and service infrastructure that hides the heterogeneity of the underlying networks and supports the multiplicity of communication needs across different types of users and user groups in emergencies.

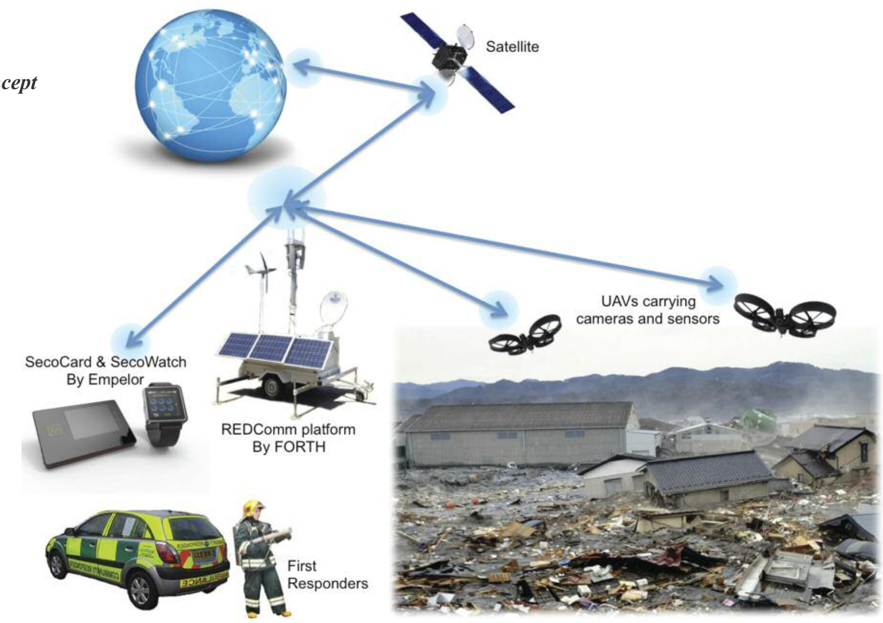
REDComm, which stands for Rapid Emergency Deployment Communication, is a trailer based communication node that utilizes several communication technologies in order to provide multiple communication services in emergency and crisis situations [2]. Such services include not only traditional communications of emergency response authorities, but also modern services to these authorities such as live video streaming and multimedia content sharing as well as public addressing and victim communication. A REDComm node includes a hybrid power source based on both renewable and non-renewable energy generators and batteries to provide electric autonomy. A pneumatic telescopic mast is installed to support communication antennas providing mobility and increased coverage range.

One or more REDComm nodes can be easily and quickly deployed anywhere to provide communication services. The usage of mesh networking forms a redundant, seamless, self-healing, backbone network that is able to route communication traffic dynamically over the most appropriate path and technology. Eight REDComm nodes have been designed and implemented by FORTH-ICS and will be evaluated in drills and real-life situations with the Hellenic emergency response authorities such as police, fire department, the national emergency aid center and the Region of Crete that participate in the REDComm project [1].

SecoCard is an intelligent - and at the same time highly secure - external token with its own screen and capacitive touch keypad, which communicates with a non-modified smartphone or tablet over Bluetooth or WiFi and runs all the security applications, while the smartphone or the tablet, respectively, just provides connectivity and their normal unshielded applications [3]. The dedicated hardware is not much larger than a few credit or key cards stacked on one another, and can operate with practically every new smartphone and tablet model being sold today or unveiled in the future by the mobile industry.

With the federation of REDComm and SecoCard, SIREN implements an integrated comprehensive security architecture based on dedicated security mechanisms, taking into

Figure 1: The SIREN concept


Satellite
UAVs carrying cameras and sensors
SecoCard & SecoWatch By Empelor
REDComm platform By FORTH
First Responders

account cross-layer considerations and multi-operator emergency environments. In terms of cryptographic functions, SIREN implements a family of key agreement methods with particular focus on password-based, weak to strong authentication associated with several multiparty contributory key agreement schemes.

SIREN is also inspired by the opportunities and threats to the security of individuals owing to the convergence of the cyber and physical world, especially in urban environments. Increasingly within such environments, the everyday life of people takes place around cyber-physical objects that are becoming smarter and more inter-connected. This creates the possibility of continual acquisition, correlation and analysis of information that can improve citizen security and safety. It is, for example, possible to combine surveillance camera information with human physiological data and incident alerts created by mobile phone users in order to establish the occurrence of incidents and reacting to them on the spot and in a personalized manner. At the same time, the closer coupling of the physical world with cyber systems creates new threats and risks for people. Access to the location of individual (or groups of) people may, for instance, trigger targeted actions against them, and mass surveillance can compromise the privacy of individuals. Risks and threats may also increase significantly if functions, critical for the security and safety of citizens, depend on a cyber-physical infrastructure and Smart City applications that are not well protected against attacks (e.g., jamming communications between emergency responders).

Enhancing these opportunities and managing the associated risks can be based on SIREN, which in the future will enable the development and runtime operation and management of different Cyber-Physical and participatory sensing applications, including aerial unmanned vehicles (drones), supporting the acquisition and sharing of security related information through the use of SIREN infrastructure. The SIREN platform will enable the development and runtime operation and management of such applications by offering an open and extensible set of integrated basic capabilities, with particular focus on emergency response and crisis management.

Links:
http://www.redcomm-project.eu
http://www.secocard.ch/

References:

[1] A. Miaoudakis et al.: "Communications in Emergency and Crisis Situations", Distributed, Ambient, and Pervasive Interactions, Springer International Publishing, 555-565, 2014.
[2] I. Askoxylakis et al.: "A Rapid Emergency Deployment Mobile Communication Node", IEEE Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD), 2014
[3] P. Papagrigoriou, et al.: "Discrete Hardware Apparatus and Method for Mobile Application and Communication Security", Human Aspects of Information Security, Privacy, and Trust. Springer International Publishing, 102-112, 2014.

Please contact:

Ioannis Askoxylakis, FORTH-ICS, Greece
E-mail: asko@ics.forth.gr

MYVISITPLANNER: Cloud-based Recommender for Personalized Tourism

by Ioannis Refanidis and Christos Emmanouilidis

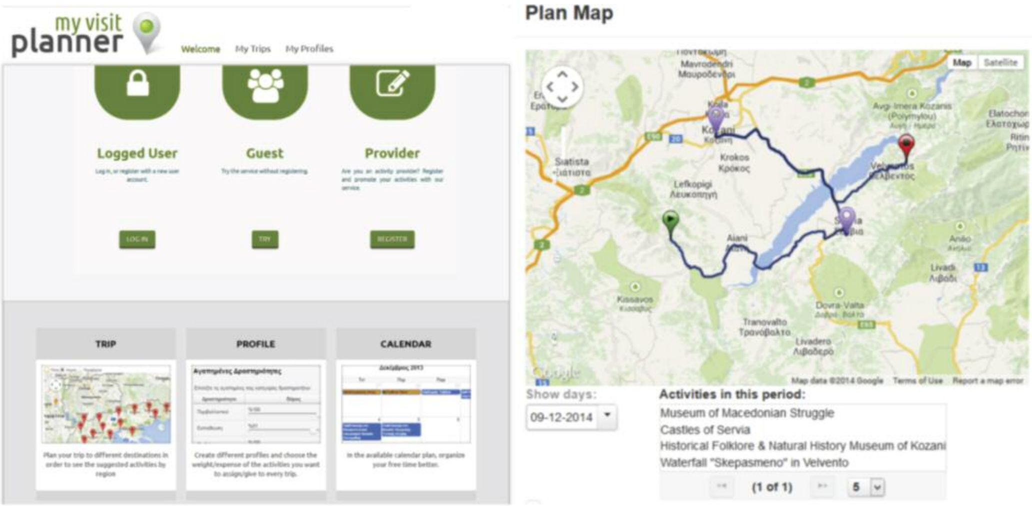
Following the masses is one way of being a tourist. But the modern creative tourist is eager to forge a personal trail. Tailoring an itinerary to individual desires has always been the realm of highly specialized tour operators. Context-aware computing and recommender systems make it possible to offer personalized services in tourism. MYVISITPLANNER is a cloud-based service employing a recommender engine to offer personalized suggestions for tour activities and a planning tool to create appropriate tour itineraries. Recommendations are offered on the basis of a hybrid approach that takes into account both a taxonomy of possible activities, as well as user preferences and past recommendations, thus offering recommendations tailored to the visit profile [1].

A creative tourist is a traveller who is not simply satisfied with visiting top-10 attractions but seeks to enjoy a personal experience when visiting a place. Whether the visit involves outdoors activities off the beaten track or certain cultural preferences, an individual traveller is often a more demanding tourist but also one that intends to better blend with local culture and people when travelling. The personalization of the tourist product can now be achieved by advanced computer-assisted tourism services.

Visitors may obtain: (i) activity recommendations contextualized by the time, duration and area of a visit, as well as by personal preferences and visit profiling; and (ii) tour itineraries created by a world-class scheduler, on the basis of the offered recommendations, while taking into account individual visitor calendar constraints and scheduling preferences, as well as available time for the visit. Activity providers can benefit from having their services enlisted and included in the itineraries recommendations (Figure 1).

While many recommender systems base their recommendations on either distance-based retrieval principles or collaborative filtering performed over past evaluations, the MYVISITPLANNER cloud service employs a hybrid recommender engine that fuses both approaches and is thus able to offer relevant recommendations even in the absence of historical data and past user evaluations feedback (Figure 2).

In the absence of past visit data, the recommendations are based on a dedicated activities ontology. Specifically, a new visit profile is mapped on the activities ontology and, based on a dedicated distance function, relevant recommendations are retrieved. Past user evaluations are handled by hybrid clustering performed over the cloud employing the Mahout cloud-oriented machine learning library. Special care is taken so that privacy-preserving data management is involved: visit profiling is preferred instead of user profiling, avoiding handling sensitive private data. The user has the



Log in, or register with a new user account.

Try the service without registering.

Are you an activity provider? Register and promote your activities with our service.

Plan your trip to different destinations in order to see the suggested activities by region

Create different profiles and choose the weight/expense of the activities you want to assign/give to every trip.

In the available calendar plan, organize your free time better.

Show days: 09-12-2014

Activities in this period:

Museum of Macedonian Struggle
Castles of Servia
Historical Folklore & Natural History Museum of Kozani
Waterfall "Skepasmeno" in Velvento

(1 of 1) 5

Figure 1: MYVISITPLANNER itinerary recommender.

option of editing the recommendations by removing activities or including activities that were not originally recommended.

In order to produce plausible plans, the scheduler takes into account both the time and space constraints imposed by the selected activities and the user's other commitments, as well as the user preferences for: (i) specific activities, (ii) the pace of the visit (e.g. relaxed or thight plan), (iii) activity duration (typical, minimum, maximum), (iv) free time scheduling, and (v) night time rest.

The scheduler produces several qualitative, significantly different alternative plans, for the user to choose from. The plan can be exported to the user's calendar.

Default options are set for users not seeking a high level of personalization and seeking to minimize the time spent on producing a plan. The service is pilot-tested with stakeholders in Northern Greece and can create whole area (as opposed to city-only) visit itineraries.

The research is conducted by a partnership between academic and research institutions (University of Macedonia, ATHENA Research and Innovation Centre) and private commercial (Gnomon Informatics) and not-for-profit (Ethnological Museum of Thrace) organizations, with support from Regional Development Agencies (Development Agency of West Macedonia), Greece.

The project is supported by the General Secretariat of Research and Technology (GSRT/MIA-RTDI), Ministry of Education and Religious Affairs,Grant no. 09SYN-62-1129, under the national Operational Program "Competitiveness and Entrepreneurship (OPC II)" (NSRF 2007-2013).

Link: http://www.myvisitplanner.com

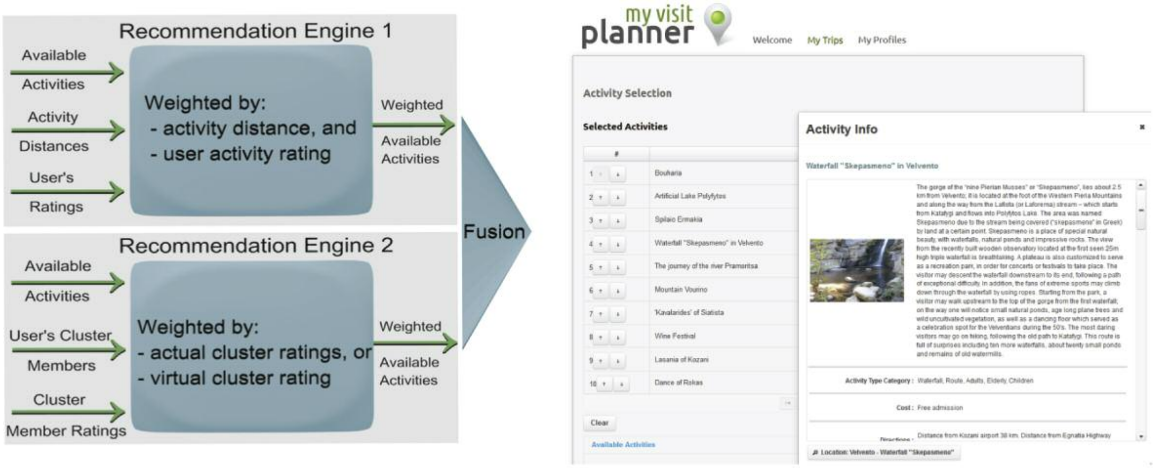
Reference:

[1] I. Refanidis et al.: "myVisitPlannerGR: Personalised itinerary planning system for tourism", in Artificial intelligence: methods and applications, A. Likas, K. Blekas, and D. Kalles, eds. (Springer), pp. 615-629.

Please contact:

Ioannis Refanidis, (Project co-ordinator)
University of Macedonia, Greece
Tel: +302310891859
E- mail: yrefanid@uom.gr

Christos Emmanouilidis
Athena Research and Innovation Centre, Greece
Tel: +302541078787
E-mail: christosem@ceti.athena-innovation.gr



Activity Selection

Selected Activities

#	
1	Bouharia
2	Artificial Lake Polyfytos
3	Spilaio Ermakia
4	Waterfall "Skepasmeno" in Velvento
5	The journey of the river Pramoritsa
6	Mountain Vourino
7	'Kavalarides' of Siatista
8	Wine Festival
9	Lasania of Kozani
10	Dance of Rokas

Activity Info

Waterfall "Skepasmeno" in Velvento

The gorge of the "nine Pierian Musses" or "Skepasmeno", lies about 2.5 km from Velvento; it is located at the foot of the Western Pieria Mountains and along the way from the Lallista (or Laforema) stream – which starts from Katafygi and flows into Polyfytos Lake. The area was named Skepasmeno due to the stream being covered ("skepasmeno" in Greek) by land at a certain point. Skepasmeno is a place of special natural beauty, with waterfalls, natural ponds and impressive rocks. The view from the recently built wooden observatory located at the first seen 25m high triple waterfall is breathtaking. A plateau is also customized to serve as a recreation park, in order for concerts or festivals to take place. The visitor may descent the waterfall downstream to its end, following a path of exceptional difficulty. In addition, the fans of extreme sports may climb down through the waterfall by using ropes. Starting from the park, a visitor may walk upstream to the top of the gorge from the first waterfall; on the way one will notice small natural ponds, age long plane trees and wild uncultivated vegetation, as well as a dancing floor which served as a celebration spot for the Velventians during the 50's. The most daring visitors may go on hiking, following the old path to Katafygi. This route is full of surprises including ten more waterfalls, about twenty small ponds and remains of old watermills.

Activity Type Category : Waterfall, Route, Adults, Elderly, Children

Cost : Free admission

Directions : Distance from Kozani airport 38 km. Distance from Egnatia Highway

Location: Velvento - Waterfall "Skepasmeno"

Figure 2: Recommendation engine principle and interface for reviewing recommendations.

Combinatorial Problem Solving for Fair Play

by Mats Carlsson

To create a fair timetable for the men's handball league is a much more complex task than you would think. There are actually many more ways to do it than there are atoms in the universe, and only one of them is perfect. The trick is to find it!

The top Swedish men's handball league, Elitserien, consists of two divisions of seven teams each. Every season, a timetable for 33 periods needs to be constructed. In the first seven periods, two parallel tournaments are played with every team meeting every other team from the same division. In the next 13 periods, one league-level tournament is played with every team meeting every other team. In the last 13 periods, the league-level tournament is played again, in reverse order.

The timetable must satisfy a number of other rules, such as:
- If team A plays team B at home, then team B must play team A at home the next time they meet.
- Teams can play at home at most twice in a row and away at most twice in a row, and such cases should be minimized.
- Both divisions must have three pairs of complementary schedules.
- Specific high-profile matches between given teams should be scheduled in specific periods.
- Some teams can't play at home during specific periods, because the venue is unavailable.

Visually, you can think of it as a matrix with 280 cells, each filled with a number between 1 and 27. There are more than 10^{400} ways to do the task, or commonly expressed 1 followed by 400 zeroes. As a comparison there are only about 10^{80} atoms in the universe. Out of the vanishingly small amount of correctly filled matrices that you will find, there is one optimal solution. The trick is to find it.

Traditionally, the time-tabling has been carried out manually by the Swedish Handball Federation. The problem is too difficult for a human to solve to optimality, and so the Federation has always had to compromise and break some rules in order to come up with an acceptable timetable. The method from SICS solves it without breaking any rules.

Researchers at KTH, the Royal Institute of Technology, had a first attempt at the problem. They conducted an initial formal study of the Elitserien schedule problem, and discovered some important structural properties. SICS continued the study, formally modelled the problem using Constraint Programming, and was thereby able to solve it to optimality in about five CPU seconds.

They first defined the variables to be used in the CP set-up, and then the essential constraints to ensure the resultant schedule will satisfy Elitserien's structural requirements. Next they highlighted some implied constraints and symmetry breaking properties that they found would greatly reduce the search effort. Finally, they modelled the league's seasonal constraints so they could construct the entire schedule in an integrated approach. The constraint model was encoded in MiniZinc 1.6 and executed with Gecode 3.7.0 as back-end.

PHoto: Shutterstock

This timetabling problem is a typical combinatorial problem. Constraint programming has been used for sports scheduling before. However, case studies solved by integrated CP approaches are scarce in the literature. Perhaps the problems have been assumed to be intractable without decomposition into simpler sub-problems.

References:

[1] J Larson, M Johansson, M Carlsson: "An Integrated Constraint Programming Approach to Scheduling Sports Leagues with Divisional and Round-Robin Tournaments", CPAIOR, LNCS 8451, pp. 144-158, Springer, 2014

[2] J. Larson, M. Johansson: "Constructing schedules for sports leagues with divisional and round-robin tournaments", Journal of Quantitative Analysis in Sports (2014), DOI:10.1515/jqas-2013-0090

Please contact:

Mats Carlsson, SICS Swedish ICT, Sweden
E-mail: matsc@sics.se

Optimizing Text Quantifiers for Multivariate Loss Functions

by Andrea Esuli and Fabrizio Sebastiani

Quantification - also known as class prior estimation – is the task of estimating the relative frequencies of classes in application scenarios in which such frequencies may change over time. This task is becoming increasingly important for the analysis of large and complex datasets. Researchers from ISTI-CNR, Pisa, are working with supervised learning methods explicitly devised with quantification in mind.

In some applications involving classification the final goal is not determining which class(es) individual unlabelled data items belong to, but determining the prevalence (or 'relative frequency') of each class in the unlabelled data. This task has come to be known as 'quantification'.

For instance, a company may want to find out how many tweets that mention product X express a favourable view of X. On the surface, this seems a standard instance of query-biased tweet sentiment classification. However, the company is likely not interested in whether a specific individual has a positive view of X but in knowing how many of those who tweet about X have a positive view of X; that is, the company is actually interested in knowing the relative frequency of the positive class.

Quantification (also known as 'class prior estimation', or 'prevalence estimation') has several applications, in fields as diverse as machine learning, sentiment analysis, natural language processing [1], data mining, social science [3], epidemiology, and resource allocation. The research community has recently shown a growing interest in tackling quantification as a task in its own right, instead of a mere byproduct of classification. One reason for this is that quantification requires evaluation measures that are different from those used for classification. Second, using a classifier optimized for classification accuracy is suboptimal when quantification accuracy is the real goal, since a classifier may optimize classification accuracy at the expense of bias. Third, quantification is predicted to be increasingly important in tomorrow's applications; the advent of big data will result in more application contexts in which analysis of data at the aggregate rather than the individual level will be the only available option.

The obvious method for dealing with quantification is to classify each unlabelled document and estimate class prevalence by counting the documents that have been attributed the class. However, when a standard learning algorithm is used, this strategy is suboptimal since, as observed above, classifier A may be more accurate than classifier B but may also exhibit more bias than B, which means that B would be a better quantifier than A.

In this work (see [2] for details) we take an 'explicit loss minimization' approach, based upon the use of classifiers explicitly optimized for the evaluation function that we use for assessing quantification accuracy. Following this route for solving quantification is non-trivial, because the measures used for evaluating quantification accuracy are inherently non-linear and multivariate, and the assumption that the evaluation measure is instead linear and univariate underlies most existing discriminative learners, which are thus suboptimal for tackling quantification.

In order to sidestep this problem we adopt the 'SVM for Multivariate Performance Measures' (SVMperf) learning algorithm proposed by Joachims, and instantiate it to optimize Kullback-Leibler Divergence, the standard measure for evaluating quantification accuracy; we dub the resulting system SVM(KLD). SVMperf is a learning algorithm of the Support Vector Machine family that can generate classifiers optimized for any non-linear, multivariate loss function that can be computed from a contingency table, such as KLD. SVMperf is a learning algorithm for 'structured prediction', i.e., an algorithm designed for predicting multivariate, structured objects. It is fundamentally different from conventional algorithms for learning classifiers: while the latter learn univariate classifiers (i.e., functions that classify individual instances independently of each other), SVMperf learns multivariate classifiers (i.e., functions that jointly label all the instances belonging to a set S). By doing so, SVMperf can optimize properties of entire sets of instances, properties (such as KLD) that cannot be expressed as linear functions of the properties of the individual instances.

Experiments conducted on 5,500 binary text quantification test sets, averaging 14,000+ documents each, have shown that SVM(KLD) outperforms existing state-of-the-art quantification algorithms both in terms of accuracy and sheer stability, and is computationally more efficient than all but the most trivial algorithms.

Link:
http://nmis.isti.cnr.it/sebastiani/Publications/TKDD15.pdf

References:

[1] Y. S. Chan, H. T. Ng: "Word sense disambiguation with distribution estimation", in proc. of IJCAI 2005, Edinburgh, UK, 1010–1015, 2005.
[2] A. Esuli, , F. Sebastiani: "Optimizing Text Quantifiers for Multivariate Loss Functions", ACM Transactions for Knowledge Discovery and Data, forthcoming.
[3] D. J. Hopkins, G.King: "A method of automated nonparametric content analysis for social science", American Journal of Political Science 54, 1, 229–247, 2010.

Please contact:

Andrea Esuli - ISTI-CNR
Tel: +39 050 3152 878
E-mail: andrea.esuli@isti.cnr.it

Dynacargo: An Urban Solid Waste Collection Information System

by Artemios G. Voyiatzis and John Gialelis

The aim of the Dynamic Cargo Routing on-the-go (Dynacargo) project is to optimize the fleet operation for the city waste collection through urban-scale sensor networks, delay-tolerant networking, citizen participation, and dynamic routing.

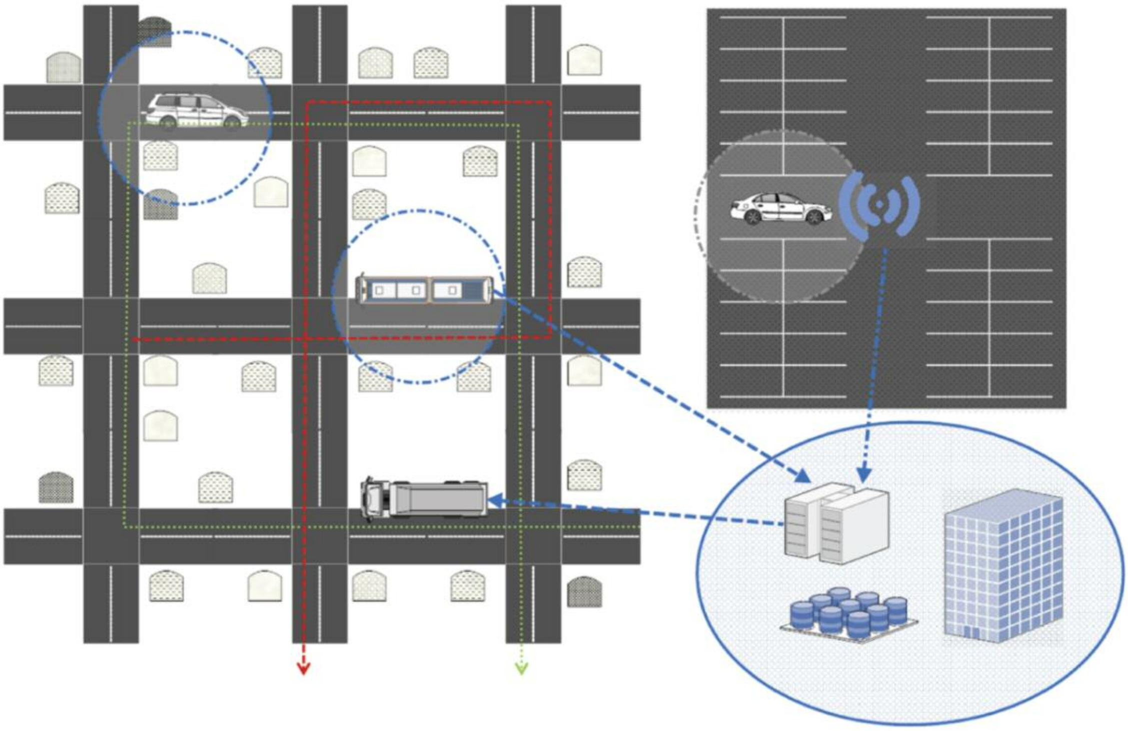
The Dynamic Cargo Routing on-the-go (Dynacargo) project aims to develop a state-of-art waste collection information system that is capable of dynamically optimizing the service plans of the garbage trucks (vehicle routes) based on the collection of waste bin fill level information. The project involves the TEI of Western Greece as a lead partner and the Industrial Systems Institute of R.C. “Athena” as scientific partners, two innovative SMEs from the ICT sector (Mindware and OrasysID), and the Municipality of Nafpaktia, Greece. The system is tested in the historical city of Nafpaktos (Lepanto). The project started in September 2013 and concludes in June 2015.

The efficient collection of urban solid waste is a growing problem for cities and urban areas. Optimized service plans can save money, reduce emissions, and improve a city’s image. However, waste bins are still emptied by experience and ‘a good guess approach’ although the advantages of utilizing ‘actual status information’ are evident [1].

In the Dynacargo project, we seek for applicable solutions at the whole chain of information collection and diffusion (Figure 1). At the bin level, we research sensors, mainly ultrasound, that cope with the harsh environment and can provide accurate fill level information while randomly-placed solid objects of different shapes are placed in the bin. Given that the bins are installed in the roads or pavements, battery operation must be assumed and strategies for longevity must be devised.

At an urban scale of operation, thousands of sensors are sparsely deployed in the city terrain. System installation and maintenance, ensuring radio coverage, and retaining network formation can become an unmanageable task. The Dynacargo project opts for low-range, point-to-point communications based on RFID technology so as to cope with these issues. Vehicles roaming around the city and equipped with readers collect the information from the bins. In order to cope with increased telecommunication costs and infrastructure upgrades, these mobile sinks defer transmissions until an Internet connection becomes available. The Delay Tolerant Networking paradigm ensures that the information is retained in the bins, until a mobile sink passes nearby and then in the sinks, until an Internet connection is available.

At the information system in the backend, the information from the mobile sinks is received and stored, along with historical data and information regarding the status of the garbage truck fleet and the city infrastructure. Artificial intelligence techniques are used to estimate missing fill level reports from the waste bins. The service plan for the fleet is modelled as a Capacitated Vehicle Routing Problem (CVRP). The Dynacargo route optimization module computes optimized service plans while the fleet is on-the-go.



Figyure 1: Dynacargo information collection and diffusion: roaming vehicles collect bin information and transmit from nearby hotspots to the backend system. Service plan (green line) of garbage truck is updated on-the-go (red line) based on processed information.

At the garbage truck, a notification system is installed in the form of a mobile app. The app assists the city workers by providing a rich-media interface to get information about the service plan and the reported or estimated waste bin fill levels. Furthermore, it provides videoconferencing services with the city operating centre based on LTE technology so as to ease notification and re-arrangements while on the go and to report any unforeseen issues on the spot (e.g. a vandalized waste bin).

The citizens interact through a web portal and a mobile app with the Dynacargo system. Individuals can learn about service plans, bin locations, and their fill level (historical, estimated, and updated information). The mobile app can also be used for crowdsourcing fill level and status information from the field, effectively engaging the citizens in the collection process and transforming them to ‘prosumers’, i.e., both producers of solid waste and information about it and also consumers of the information regarding the waste collection city service.

Currently, the Dynacargo project has collected the requirements and designed the system architecture [2,3]. The various system components are finalized and the system integration process initiated. In the next few months, a small-scale prototype of the system will be installed in the city of Nafpaktos, Greece so as to test and demonstrate its operation in a real environment.

Efficient solid waste collection from cities and urban areas is a priority issue for city authorities. The benefits of a streamlined operation based on ICT are apparent for the smart city of the future and its stakeholders, including the city authorities, the inhabitants, the environment, and the quality of experience for the visitors.

Link:
http://dynacargo.cied.teiwest.gr/

References:

[1] M. Faccio et al.: “Waste collection multi objective model with real time traceability data”, Waste Management 31, 2011.

[2] G. Asimakopoulos et al.: “Architectural modules of a dynamic collection management system of urban solid waste”, in proc. of ITS 2014: ITS and Smart Cities, 2014.

[3] A.G. Voyiatzis et al.: “Dynamic cargo routing on-the-go: the case of urban solid waste collection”, 2nd IEEE WiMob 2014 International Workshop on Smart City and Ubiquitous Computing Applications (SCUCA 2014), 2014.

Please contact:

John Gialelis
Industrial Systems Institute, “Athena” RIC in ICT and Knowledge Technologies, Greece
E-mail: gialelis@isi.gr

Joint Collaborative Workshops of the ERCIM Dependable Embedded Software-intensive Systems Working Group

by Erwin Schoitsch

The ERCIM Dependable Embedded Software-intensive Systems Working Group (DES WG) organized a collaborative Workshop at SAFECOMP 2014, the International Conference on Computer Safety, Reliability and Security and a special session at at the joint Euromicro Conference on Digital System Design and Software Engineering and Advanced Applications (DSD/SEAA 2014) . The events were organized jointly with the European Workshop on Industrial Computer Systems, Technical Committee 7 (EWICS TC7) and projects of the European Technology Platform and Initiative ARTEMIS (Advanced Research and Technology for Embedded Intelligence and Systems).

SAFECOMP DECSoS Workshop:

The Workshop on Dependable Embedded and Cyber-physical Systems and Systems-of-Systems (DECSoS’14) was organized as one of the co-located workshops at SAFECOMP 2014, 9 September 2014 in Florence, Italy. Erwin Schoitsch (AIT, Austria) and Amund Skavhaug (NTNU, Norway) were largely responsible for its organization, together with an international program committee composed of 15 experts from the ERCIM, EWICS TC7 and SAFECOMP organizers, who reviewed the papers. The workshop was co-hosted by the ARTEMIS projects MBAT, CRYSTAL, EMC², SafeCer and ARROWHEAD. The presented papers are published by Springer in the LNCS series (LNCS 8696), the abstracts are available on the Springer Web site .

The workshop DECSoS’14 comprised the sessions Formal Analysis and Verification, Railway applications: Safety analysis and verification, and Resilience and Trust: Dynamic issues.

About 30 participants attended the workshop. The mixture of topics was well balanced, with a focus on Software and System Analysis and Verification, and addressed by the ARTEMIS projects MBAT, CRYSTAL and EMC², supporting the goal of collaboration and experience exchange between related ARTEMIS projects. These projects are building on each other’s results and achievements, working towards a common Collaborative Reference Technology Platform (CRTP) based on an IOS Specification (Interoperability Specification). The aim is to build a sustainable innovation eco-system around the “High-Reliability-Cluster” of ARTEMIS projects. Overall, the workshop provided interesting insights into the topics, and enabled fruitful discussions both during the meeting and afterwards.

The annual SAFECOMP conference is a leading conference in this area, focusing on industrial computer control systems and applications. Since it was established in 1979 by the European Workshop on Industrial Computer Systems, Technical Committee 7 on Reliability, Safety and Security (EWICS TC7), SAFECOMP has contributed to the progress of the state-of-the-art in dependable application of computers in safety-related and safety-critical systems.

SAFECOMP 2014 was the 33rd International Conference on Computer Safety, Reliability and Security, and took place in Florence, Italy, from Sept. 8-12, 2014.

SAFECOMP covers state-of-the-art, experience and new trends in the areas of safety, security and reliability of critical computer applications. SAFECOMP provides ample opportunity to exchange insights and experience on emerging methods, approaches and practical solutions.

ERCIM/ARTEMIS/EUROMICRO Special Session TET-DEC

A joint Special Session TET-DEC "Teaching, Education and Training for Dependable Embedded and Cyber-physical Systems" was held at the Euromicro Conference on Digital System Design (DSD)and Software Engineering and Advanced Applications (SEAA) 2014 in Verona, Italy, 27-29 August 2014. It was jointly organized by the ERCIM DES WG, Euromicro and the ARTEMIS Education & Training Working Group.

In the field of Cyber-physical Systems and Systems of Systems, there is tremendous investment in research and innovation. Gaps still exist, however, in education and training in this area. After a first successful start of this session in 2013, we asked again this year: "How should we educate and train our current and future engineers and researchers? This special workshop show-cased the current work in this area, facilitating fruitful discussions and exchanges of ideas, presenting best practices and experience reports, and analysis of the challenges with recommendations for a sustainable future.

The workshop of the ERCIM DES WG was supported by the ARTEMIS E&T (Education & Training) Working Group, ARTEMIS projects, in particular SafeCer ("Safety Certification of Software-Intensive Systems with Reusable Components"), as well as MBAT (Combined Model-based Analysis and Testing of Embedded Systems) and R3-COP (Resilient Reasoning Robotic Co-operating Systems).

The TET-DEC workshop was part of the regular session of the conference and the papers are in the IEEE conference proceedings. The special session included five presentations. It also provided an overview on support for education and training activities in European and national research projects in the area of embedded systems:

- Erwin Schoitsch, Introduction: Teaching, Education and Training viewed from European projects' perspectives.
- Miren Illarramendi Rezabal, Leire Etxeberria Elorza and Xabier Elkorobarrutia Letona. Reuse in Safety Critical Systems: Educational Use Case First Experiences.
- Jakob Axelsson, Avenir Kobetski, Ze Ni, Shuzhou Zhang and Eilert Johansson. MOPED: A Mobile Open Platform for Experimental Design of Cyber-Physical Systems (one of three papers to receive the "Best Paper Award" of the conference)
- Elena Gomez-Martinez and Alvaro Fernandez-Diaz. Introducing Embedded Systems to People with Special Needs: Insights from a Real Case
- Clara Benac Earle, Lars-Ake Fredlund, Julio Marino and Thomas Arts. Teaching students Property-based Testing.

Links:
SAFECOMP 2014: http://www.safecomp2014.unifi.it/
SAFECOMP 2014 Workshop proceedings:
http://link.springer.com/book/10.1007/978-3-319-10557-4
Euromicro: http://www.euromico.org
ARTEMIS projects: http://www.artemis-ia.eu/all-projects.html

Please contact:
Erwin Schoitsch, AIT Austrian Institute of Technology
E-mail: erwin.schoitsch@ait.ac.at

MUSCLE Working Group International Workshop on Computational Intelligence for Multimedia Understanding

by Maria Trocan, Emanuele Salerno and Enis Cetin

The Institut Superieur d'Electronique de Paris (ISEP) hosted the International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM 2014), organized by the ERCIM Working Group on Multimedia Understanding through Semantics, Computation and Learning (Muscle), 1-2 November 2014.

Multimedia understanding is an important part of many intelligent applications in our social life, be it in our households, or in commercial, industrial, service, and scientific environments. Analyzing raw data to provide them with semantics is essential to exploit their full potential and help us in managing our everyday tasks. The purpose of the workshop was to provide an international forum to present and discuss current trends and future directions in computational intelligence for multimedia understanding. The workshop also aimed at fostering the creation of a permanent network of scientists and practitioners for easy and immediate access to people, data and ideas. This is now the third such workshop organized by MUSCLE. As in the past, the participation was open to all interested researchers. This year, the papers presented, as well as the audience, were particularly numerous, thus strengthening the networking capabilities of the group with some more research teams requesting to join MUSCLE.

34 participants from eleven countries attended the workshop. The papers accepted for presentation and publication in IEEE Xplore were 27. The presentations were divided into three additional thematic sessions, with respect to traditional tracks: Big and linked data, Hyperspectral image processing,

and Retinal image processing and analysis. The talks covered a very wide range of subjects. A motive for satisfaction was the presence of several papers dealing with media different from image and video, some of them integrating multiple media to approach understanding. Three authoritative invited speakers presented keynote talks: Prof. Gauthier Lafruit, of Université Libre de Bruxelles (Image-based 3D scene visualization and Free Viewpoint TV), Prof. Michel Crucianu, of the Conservatoire National des Arts et Métiers, Paris (Multimedia Information Retrieval: Beyond Ranking), and Prof. François-Xavier Coudoux, of the Institute of Electronics, Microelectronics, and Nanotechnologies, France (Digital Image and Video Transcoding: Application to Optimized Video Delivery over Error-prone Networks).

The Muscle group has more than 60 members from 23 partner groups in 16 countries. Their expertise ranges from machine learning and artificial intelligence to statistics, signal processing and multimedia database management.

Links:
http://iwcim.isep.fr/index.html
http://wiki.ercim.eu/wg/MUSCLE/

Please contact:
Maria Trocan, ISEP, and Emanuele Salerno, ISTI-CNR, General Chairs, IWCIM 2014
A. Enis Cetin, Bilkent University, Technical Program Committee Chair, IWCIM 2014
E-mail maria.trocan@isep.fr, emanuele.salerno@isti.cnr.it

Networking: IFIP TC6 2014 Dagstuhl Meeting

by Harry Rudin

The International Federation for Information Processing's (IFIP's) Technical Committee 6 (TC6) held its biannual meeting at Schloss Dagstuhl, Germany, on 2- 14 November 2014.

TC6 concerns itself with Communications Systems. The objective was to start designing a strategic plan for TC6's future activities. To this end several talks were given of general interest, many dealing with communication networking, both in the measurement and international cooperation sense. Much of this information is of wide interest and publicly accessible. This broad interest is the reason for this report. The overall Dagstuhl program is available at http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=14463

Vassilis Kostakos from the University of Oulu, Finland presented his panOULU system which uses wireless mobility traces for mobility analysis and traffic planning in Oulu. At http://ufn.virtues.fi/~swproject12/ one can few current hotspots of activity in Oulu.

Alessandro D'Alconzo from the Telecommunication Research Center in Vienna talked about his mPlane system, a distributed measurement infrastructure to perform active, passive and hybrid measurements of the dynamics of the Internet (http://www.ict-mplane.eu/).

Panayotis Antoniadis form the ETH in Zurich discussed his NetHood Initiative. The project has, among other work, produced a small, inexpensive transceiver for establishing an adhoc network in a small community. The group is looking for suggestions for additional applications of these devices. See http://nethood.org/first_draft.html

Fabio Ricciato from the Austrian Institute of Technology in Vienna described the TMA (Traffic Monitoring and Analysis) portal. TMA was originally supported by COST11 but now has a life of its own for coordinating network measurements (http://www.tma-portal.eu/)

Georg Carle from the Technical University in Munich described the Network of Excellence in Internet Science (EINS). EINS facilitates cooperation among researchers on an international basis (http://www.internet-science.eu/)

Filip De Turck, University of Ghent, described the clustering effort in Europe to coordinate ongoing European research on the Internet of Things (http://cordis.europa.eu/fp7/ict/enet/rfid-iot_en.html).

IFIP digital libraries were also discussed. TC6 now has its own Web site up and working (http://opendl.ifip-tc6.org). One possible addition discussed was a TC6 networking journal, published only online and relatively rarely with the feature of extraordinarily strong articles. One thought discussed was reimbursing reviewers so as to obtain equally strong and constructive reviews. Clearly the journal would be open access.

Another discussion was on the quality and ranking of communication conferences. Obviously this is an important issue inside and outside the IFIP community, given today's demand for research publication to strengthen the resume. The purpose of these discussions was to consider various possibilities for IFIP's TC6 to be of greater service to the networking community. Extending our digital library, supporting COST or Network of Excellence projects under an IFIP umbrella, providing conference ranking, and extending our educational reach are all under consideration.

Links:
http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=14463
http://ufn.virtues.fi/~swproject12
http://www.ict-mplane.eu/
http://www.tma-portal.eu/
http://www.internet-science.eu/
http://cordis.europa.eu/fp7/ict/enet/rfid-iot_en.html
http://opendl.ifip-tc6.org

Please contact:
Harry Rudin, Swiss Representative to IFIP TC6
E-mail: hrudin@sunrise.ch

Call for Papers

Teaching, Education and Training for Dependable Embedded Systems at SEAA/DSD 2015

Funchal, Madeira, Portugal, 26-28 August 2015

A special session on "Teaching, Education and Training for Dependable Embedded and Cyber-physical Systems"(TET-DEC), co-organised by the ERCIM Dependable Embedded Systems Working Group, will be held at the Euromicro Conference on Digital System Design and Software Engineering and Advanced Applications (DSD/SEAA 2015).

In embedded systems and cyber-physical systems research, there is tremendous investment in research and innovation – but is this complemented in education and training as well? Particular issues of safety, security, dependability, risk, hazard and vulnerability analysis have to be properly trained and be part of basic university education as well as of "life-long learning" for professionals. Obviously there are gaps. Therefore, we will ask how we should educate and train our current and future engineers and researchers. This special workshop solicits reports on on-going work aiming at fruitful discussions and exchange of ideas, to present best practices, examples and experience, and analysis of the challenges with recommendations for a sustainable future.

Deadlines:
- Abstract submission: 15 Feb. 2015
- Paper submission: 22 Feb. 2015
- Notification of acceptance: 27 April 2015
- Camera-ready paper: 17 May 2015

More information:
http://paginas.fe.up.pt/~dsd-seaa-2015/seaa2015/call-for-papers-seaa-2015/tet-dec-special-session/
http://paginas.fe.up.pt/~dsd-seaa-2015/
or contact the workshop and programme committee chairpersons Erwin Schoitsch, AIT Austrian Institute of Technology and Amund Skavhaug, NTNU

Call for Papers

ERCIM/EWICS/ARTEMIS Workshop at SAFECOMP 2015

Delft, The Netherlands,
22 September 2015

SAFECOMP is an annual event covering the experience and new trends in the areas of safety, security and reliability of critical computer applications. It provides ample opportunity to exchange insights and experience on emerging methods, approaches and practical solutions. The 34th edition of SAFECOMP focuses on the challenges arising from networked multi-actor systems for delivery of mission-critical services, including issues of the "Systems-of-Systems" area, medical technology and patient care.

The already well-established ERCIM/EWICS/ARTEMIS Workshop on Dependable Embedded Cyber-physical Systems and Systems-of-Systems" co-organised by the ERCIM DES-Working Group, co-hosted by ARTEMIS projects EMC², ARROWHEAD and CRYSTAL, is again planned for the Workshop-Day 22 September 2015. For the workshop papers, a separate call will be published later, but proposals can already be sent now to erwin.schoitsch@ait.ac.at

Important dates:
- Workshop proposal submission: 16 February 2015
- Abstract submission: 26 February 2015
- Full paper submission: 2 March 2015
- Notification of acceptance: 1 2 May 2015
- Camera-ready submission: 15 June 2015

For more information, see
http://safecomp2015.tudelft.nl/
or contact erwin.schoitsch@ait.ac.at

Call for Contribution

ASQT 2015 - 13th User Symposium on Software Quality, Test and Innovation

Graz, Austria, 16-17 April 2015

The User Symposium on Software Quality, Test and Innovation is the 13th edition in a series of highly successful two-day user symposiums in the field of software testing, quality and innovation, this year co-located with ICST 2015, the 8th IEEE International Conference on Software Testing, Verification and Validation.

ASQT aims to
- facilitate exchange of industry case studies in the field of software development and operating / sourcing software,
- enable academic researchers and industry innovators to exchange ideas and results, and
- disseminate scientific results as well as expertise in the field of quality assurance and software testing

Major topics of interest include, but are not limited to: Agile and Lean Practices, Business Alignment in IT, Data Privacy and Security, Design for Testability, Industrial challenges and experience reports, Innovation and Business Models, Penetration and Security Testing, Quality Management and Key Performance Indicators, Quality of mobile and ubiquitous apps, Software as a Service, Software Testing as a Service and Managed Testing Services, Testing Methods, Tools in Software Testing, and Test Process Improvement.

Important dates:
- Submission Deadline: 8 Feb. 2015
- Notification: 1 March 2015

ASQT 2015 and ICST 2015 offer shared keynotes and a shared session (practitioner's topics). Accepted contributions to ASQT 2015 are published in the IEEE Workshop proceedings of ICST 2015.

More Information:
http://www.asqt.org
http://icst2015.ist.tugraz.at/

Antoine Petit new Chairman and CEO of Inria

© Inria / Photo C. Helsly

Antoine Petit was appointed Chairman and CEO of Inria for a term of five years on 26 September 2014. Associate Professor of Mathematics and Doctor of Science, University Professor at ENS Cachan, he was formerly Deputy Managing Director of Inria.

Antoine Petit joined Inria in July 2006 to head the Paris-Rocquencourt Research Centre, later acting as interim Director of the Saclay Research Centre from March to September 2010. As Deputy Managing Director, he coordinated relations with supervisory ministries and institutional partnerships with research bodies, businesses and local authorities, in France and abroad. He has also been in charge of training through research programmes. Antoine Petit specialises in formal methods, mainly transition system based methods, for the specification and verification of parallel and real-time systems. Antoine Petit succees Michel Cosnard who remains President of ERCIM EEIG until the end of 2014.

Manfred Hauswirth new Executive Director of Fraunhofer FOKUS

Matthias Heyde/ Fraunhofer FOKUS

Prof. Dr. Manfred Hauswirth was appointed as Director of the Fraunhofer Institute for Open Communications Systems FOKUS in Berlin. Hauswirth assumed his new post on 1 October 2014. Concurrently, he will take over the professorship in “distributed open systems” at the Technical University of Berlin. Hauswirt succeeds Prof. Dr. Popescu-Zeletin who retired from his position as Executive Director of Fraunhofer FOKUS on 30 September 2014.

The 2014 Nobel Prize in Physiology or Medicine for two Professors at NTNU

The Norwegian University of Science and Technology (NTNU) is extremely happy and proud to now have two Nobel Prize Laureates in their scientific staff. Professors May-Britt Moser and Edvard Moser at NTNU have been awarded the Nobel Prize for 2014 in Physiology or Medicine for their discoveries of how the brain creates a map of the space around us to navigate complex surroundings. They share the award with John O'Keefe of University College London, and they are only the second married couple to win a Nobel in medicine.

Professors May-Britt Moser and Edvard Moser at the Norwegian University of Science and Technology have been awarded the Nobel Prize for 2014 in Physiology or Medicine. Photo: Geir Morgen.

Their research and discoveries has been conducted at NTNU since 1996. Their work and excellence has received a lot of international recognition and prizes, and attracted funding and top class researchers from all over the world. The research results have been achieved through being focused with a clear basic research goal, and combining hard work with excellent ideas. May-Britt Moser was asked by a journalist the announcement what this can be used for, “this is basic research, we don’t know that yet, only that is definitely important to understand”. It’s a lot of wisdom in that. The Mosers appreciate the support and freedom at NTNU, and have announced that they have a lot of new ideas and unfinished work, and they will continue their research at NTNU. The award has received international and national attention, has definitely put the ERCIM member NTNU in Trondheim on the map as a university of excellence.

More information: http://www.ntnu.edu/nobelprize

New Italian Research on Monitoring Monuments and Quality of Sleep

The Fondazione Cassa di Risparmi di Lucca is Co-funding two Projects Proposed by two Institutes of the National Research Council in Pisa. The first project, MONSTER (Monito-raggio strutturale di edifici storici con tecnologie wireless e strumenti di calcolo innovativi), proposed by the Institute of Information Science and Technologies (ISTI), will monitor the Bell Tower of the San Frediano Basilica in Lucca through a non-invasive wireless sensor network (WSN). The sensors will collect data relevant to the atmospheric conditions and the stresses the tower is subjected to. The goal is a continuous, cheap and non-invasive monitoring of an historic heritage, in order to have the possibility to immediately intervene when structural problems are detected.The second project, Well Being @ Lucca (WB@ Lucca), proposed by the Institute of Informatics and Telematics (IIT), is addressed to elderly people; some parameters related to their health and the quality of their sleeping will be monitored. They will also be provided with a tool to stimulate their domestic physical activity and their smart phone will be used to increase their social contacts.

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/

Czech Research Consortium
for Informatics and Mathematics
FI MU, Botanicka 68a, CZ-602 00 Brno, Czech Republic
http://www.utia.cas.cz/CRCIM/home.html

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/

FWO
Egmontstraat 5
B-1000 Brussels, Belgium
http://www.fwo.be/

F.R.S.-FNRS
rue d'Egmont 5
B-1000 Brussels, Belgium
http://www.fnrs.be/

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/

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

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

Spanish Research Consortium for Informatics and Mathematics
D3301, Facultad de Informática, Universidad Politécnica de Madrid
28660 Boadilla del Monte, Madrid, Spain,
http://www.sparcim.es/

Science and Technology Facilities Council
Rutherford Appleton Laboratory
Chilton, Didcot, Oxfordshire OX11 0QX, United Kingdom
http://www.scitech.ac.uk/

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/

University of Geneva
Centre Universitaire d'Informatique
Battelle Bat. A, 7 rte de Drize, CH-1227 Carouge
http://cui.unige.ch

University of Southampton
University Road
Southampton SO17 1BJ, United Kingdom
http://www.southampton.ac.uk/

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

Universty of Wroclaw
Institute of Computer Science
Joliot-Curie 15, 50–383 Wroclaw, Poland
http://www.ii.uni.wroc.pl/

Technical Research Centre of Finland
PO Box 1000
FIN-02044 VTT, Finland
http://www.vtt.fi/

Subscribe to ERCIM News and order back copies at http://ercim-news.ercim.eu/