Number 89, April 2012

ERCIM NEWS

www.ercim.eu

Special theme:

Big Data

Also in this issue:

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 8,500 copies. The printed version of ERCIM News has a production cost of €8 per copy. Subscription is currently available free of charge.

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

Cover image:
A heavy ion collision event animation from the Large Ion Collider Experiment (ALICE) © CERN

Keynote

E-Infrastructures for Big Data: Opportunities and Challenges

The management of extremely large and growing volumes of data has since many years been a challenge for the large scientific facilities located in Europe such as CERN or ESA, without clear long term solutions. The problem will become even more acute as new ESFRI facilities come on-stream in the near future. The advent of "big data science", however, is not limited to large facilities or to some fields of science. Big data science emerges as a new paradigm for scientific discovery that reflects the increasing value of observational, experimental and computer-generated data in virtually all domains, from physics to the humanities and social sciences.

The volume of information produced by the "data factories" is a problem for sustainable access and preservation, but it is not the only problem. Diversity of data, formats, metadata, semantics, access rights and associated computing and software tools for simulation and visualization add to the complexity and scale of the challenge.

Big Data and e-Science: challenges and opportunities

ICT empowers science by making possible massive interdisciplinary collaboration between people and computers, on a global scale. The capacity and know-how to compute and simulate, to extract meaning out of vast data quantities and to access scientific resources are central in this new way of co-creating knowledge. Making efficient use of scientific data is a critical issue in this new paradigm and has to be tackled in different dimensions: creation of data, access and preservation for re-use, interoperability to allow cross-disciplinary exploration and efficient computation, intellectual property, etc.

ICT infrastructures for scientific data are increasingly being developed world-wide. However, many barriers still exist across countries and disciplines making interoperability and sustainability difficult to achieve. To cope with the extremely large or complex datasets generated and used in research, it is essential to take a global approach to interoperability and discoverability of scientific information resources. International cooperation to achieve joint governance, compatible legal frameworks and coordinated funding is also necessary.

Data-intensive science needs be reproducible and therefore requires that all research inputs and outcomes are made available to researchers. Open access to scholarly papers, trusted and secure access to data resources and associated software codes, and interlinking of resources with publications, they all support reproducible and verifiable e-science. In some areas the storage and processing of large datasets may have implications to data protection, which need to be investigated together with access to data by the public.

Kostas Glinos
European Commission, DG Information Society and Media
Head of GEANT and e- Infrastructure Unit

In all fields of science we can encounter similar technical problems when using extremely large and heterogeneous datasets. Data may have different structures or may not be well structured at all. Analytical tools to extract meaningful information from the huge amounts of data being produced are lagging. Technical problems are often more complex in interdisciplinary research which is the research paying the highest rewards. When the amounts of data to be processed are large they cannot easily move around the network. Novel solutions are therefore needed; and in some cases, storage and data analysis resources might need to move to where data is produced.

A significant part of the global effort should focus on increasing trust (eg through international certification) and enhancing interoperability so that data can be more readily shared across borders and disciplines. Second, we need to develop new tools that can create meaningful, high quality analytical results from large distributed data sets. These tools and techniques are also needed to select the data that is most valuable for future analysis and storage. This is a third focus of effort: financial and environmental systainability. The rate of global data production per year has already exceeded the rate of increase in global data storage capacity; this gap is widening all the time, making it increasingly more important to understand what data has an intrinsic value that should not be lost and what data is "transient" and we could eventually throw away [Richard Baraniuk, More is Less: Signal Processing and the Data Deluge Science 2011 (331): at p. 717].

European Commission activities in scientific data

Through the 7th Framework Programme for research, the Commission, in coordination with Member States, promotes and funds ICT infrastructures for research (e-infrastructures) enabling the transition to e-science. The Commission has invested more than 100 M€ in the scientific data infrastructure over the last few years, covering domains ranging from geospatial information and seismology to genomics, biodiversity and linguistics . The development of e-Infrastructures is part of the Digital Agenda flagship initiative, envisioned as means to connect researchers, instruments, data and computation resources throughout Europe. Furthermore, the 2009 Communication of the Commission on ICT infrastructures for e-science highlighted the strategic role of IT in the scientific discovery process and sought to increase adoption of ICT in all phases of this process. The Communication expressed the urgency to develop a coherent strategy to overcome the fragmentation in infrastructures and to enable research communities to better manage, use, share and preserve data. In its conclusions of December 2009, the Competitiveness Council of the European Union invited Member States and the Commission to broaden access to scientific data and open repositories and ensure coherent approaches to data access and curation.

More recently, in October 2010, the High Level Expert Group on Scientific Data submitted its final report to the Commission. The main conclusion of the report is that there is a need for a "collaborative data infrastructure" for science in Europe and globally. The vision this infrastructure would enable is described in the following terms:

"Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure a valuable asset, on which science, technology, the economy and society can advance."

A complementary vision was developed by the Commission co-funded project GRDI2020. It envisions a Research Data Infrastructure that enables integration between data management systems, digital data libraries, research libraries, data collections, data tools and communities of research.

These efforts are expected to create a seamless knowledge territory or "online European Research Area" where knowledge and technology move freely thanks to digital means. Furthermore, it is essential to take a global approach to promote interoperability, discoverability and mutual access of scientific information resources.

Financial support for this policy is expected to come from the next framework programme for research and innovation. The Commission has included data e-infrastructure as a priority in its proposals for the so-called Horizon 2020 programme, covering the period from 2014 to 2020. Coordination with funding sources and policy initiatives in Member States of the EU is also necessary as much of the e-infrastructure in Europe obtains financing and responds to needs at national level.

In summary, data should become an invisible and trusted e-infrastructure that enables the progress of science and technology. Beyond technical hurdles, this requires a European (and global) research communication system that enables and encourages a culture of sharing and open science, ensures long-term preservation of scientific information, and that is financially and environmentally sustainable.

The views expressed are those of the author and do not necessarily represent the official view of the European Commission on the subject.

RESEARCH AND INNOVATION

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

EVENTS, BOOKS, IN BRIEF

NanoICT: A New Challenge for ICT

by Mario D'Acunto, Antonio Benassi, Ovidio Salvetti

Nanotechnology is the manipulation or self-assembly of individual atoms, molecules or molecular clusters into structures to create material and devices through an exact control of size and form in the nanometer scale. The immense potential of this field is presenting a challenge for the ICT world.

A nanometer (nm) is a billionth of a meter ($1nm=10^{-9}m$), ie about 1/80,000 of the diameter of a human hair, or 10 times the diameter of a hydrogen atom. The term nanotechnology is generally used when referring to materials of size 0.1nm to 100nm. Materials with nanometric structures often exhibit quite different properties- mechanical, optical, chemical, magnetic or electronic - compared with traditional bulk materials made from the same chemical composition. Two principal factors cause the properties of nanomaterials to differ significantly from other materials: increased relative surface area, and quantum effects. These factors can change or enhance properties such as reactivity, strength and electricity, or optical characteristics, because the deviation of surface and interface properties from the bulk properties of larger amounts of material sometimes leads to unexpected surface effects.

Information and Communications Technology (ICT) is one of the areas that has most benefited from nanotechnologies, where it has been traditionally associated with nanoelectronics, in the efficient development and miniaturization of items such as computer chips, information storage, and sensors. Certain ICT procedures, such as distributed calculus and smart processing can be considered suitable for the implementation of bottom-up nanotechnology procedures. The self-assembly of nanostructures is the clearest evidence of a bottom-up processing (as opposed to miniaturization that can be considered the basic top-down procedure). Self-assembly is the art of building by mixing. Chemists have been doing this for centuries. The challenge today is to make such systems smart. In order to successfully use self-assembly to build micro- and nano-devices it is important to use building blocks that can be programmed to assemble in certain, pre-determined ways. If all the components that are being assembled are of the same kind, a simple over-scale structure will be the result. Using building blocks with differentiated binding sites that only fit together in certain patterns is a way to program the assembly process and makes it possible to build far more advanced structures than just periodic ones. Such processes are known as programmable self-assembly.

One important example of programmable self-assembly is the mechanism of single stranded DNA hybridization on different surfaces, gold nanoparticles surfaces, carbon nanotubes, etc. DNA is an excellent self-assembly glue since it is specific in its bonding interactions and can be used in various nanotechnology applications, see Figure 1. Recently, we showed that the hybridization of DNA on a single-wall carbon nanotube (SWNT) is accomplished by a band-gap fluorescent shift due to changes in the exciton population (M. D'Acunto, S. Colantonio, D. Moroni, O. Salvetti, Journal of Modern Optics, 57, 1695-1699, 2010). An exciton is a bond state between an electron crossed on the conduction band and its corresponding hole in the valence band connected by the electrostatic Coulomb force. The possibility to tune the exciton population during the self-assembling process opens the road for the production of smart biosensors with many possible applications (genetic diagnosis, screening of genetically modified food, etc.). The smart biosensors, engineered by our simulations, will implement DNA sequences that are complementary to the carbon nanotubes and are compatible with specific biosensor enzymes for many different compounds. In turn, it will be a self-assembling guided procedure for biosensors at the biomolecular level. Such smart sensors could be strongly improved using graphene sheets instead of carbon nanotubes, because of the large area for single strand DNA functionalization made available by graphene sheets, see Figure 2. The construction of a smart nano-biosensor based on the self-assembly of DNA on graphene sheets is a future exciting challenge for scientists bridging the gap between the behavior of matter on the nanoscale and the ICT world. Another possible application that we are studying is to use electrospun nanofiber for hybridizing single stranded DNA. Electrospinning is a well-developed process used to produce nanofibers from a variety of materials. In electrospinning, a

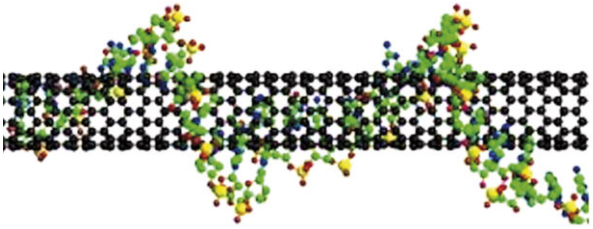

Figure 1. Example of DNA wrapping a carbon nanotube. The chemical bonding of nanotubes is composed entirely of sp2-bonds, similar to those of graphite. These bonds provide nanotubes with their unique strength, and other specific physical-chemical properties, selectively changed by the bonding with DNA filament.

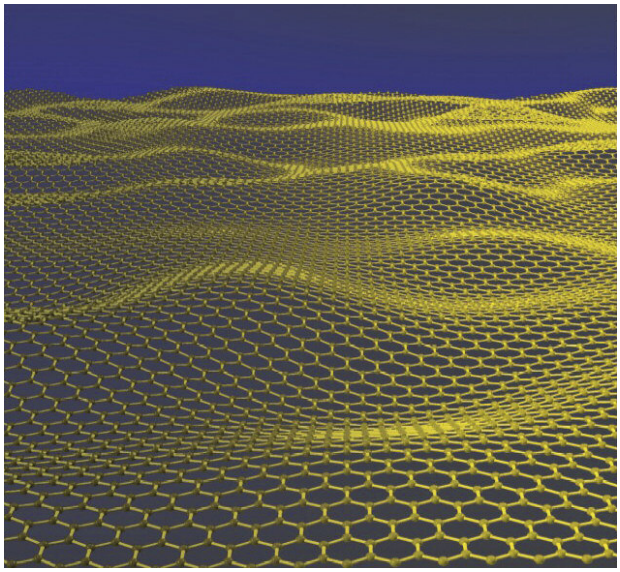

Figure 2: A graphene sheet: a one-atom-thick planar sheet of sp^2-bonded carbon atoms densely packed in a honeycomb crystal lattice. The crystalline or flake form of graphite consists of many graphene sheets stacked together.

high voltage is applied to viscous solution on a sharp conducting tip, causing it to form a Taylor cone. As the electric field is increased, a fluid jet is extracted from the Taylor cone and accelerated toward a grounded collecting substrate. Nanofibers (having at least one dimension of 100 nanometer (nm) or less) exhibit special properties mainly due to extremely high surface to weight ratio. Within this main nanoICT challenge, we are now working on simulating possible improvements of DNA-functionalized nanofibers to be used as smart nanobiosensors through a combination of selective DNA chemical bonding with nanofiber surface.

Please contact:
Mario D'Acunto, ISM-CNR, Italy
E-mail: mario.dacunto@ism.cnr.it

Ovidio Salvetti or Antonio Benassi, ISTI-CNR, Italy
E-mail: antonio.benassi@isti.cnr.it,
ovidio.salvetti@isti.cnr.it

Information Extraction from Presentation-Oriented Documents

by Massimo Ruffolo and Ermelinda Oro

The Web is the largest knowledge repository ever. In recent years there has been considerable interest in languages and approaches providing structured (eg XML) and semantic (eg Semantic Web) representation of Web content. However, most of the information available is still accessed via Web pages in HTML and documents in PDF, both of which have internal encoding conceived to present content on screen to human users. This makes automatic information extraction problematic.

In Presentation-Oriented Documents (PODs) content is laid out to provide visual patterns that help human readers to make sense of it. A human reader is able to look at an arbitrary document and intuitively recognize its logical structure and understand the various layout conventions and complex visual patterns that have been used in its presentation. This aspect is particularly evident, for instance, in Deep Web pages where Web designers arrange data records and data items with visual regularity, and in PDF documents where tables are used to meet the reading habits of humans. However, the internal representations of PODs are often very intricate and not expressive enough to allow the associated meaning to be extracted, even though it is clearly evidenced by the presentation.

In order to extract data from such documents, for purposes such as information extraction, it is necessary to consider their internal representation structures as well as the spatial relationships between presented elements. Typical problems that must be addressed, especially in the case of PDF documents, are incurred by the separation between document structure and spatial layout. Layout is important as it often indicates the semantics of data items corresponding to complex structures that are conceptually difficult to query, eg in western languages, the meaning of a cell entry in a table is most easily defined by the leftmost cell of the same row and the topmost cell of the same column. Even when the internal encoding provides fine-grained annotation, the conceptual gap between the low level representation of PODs and the semantics of the elements is extremely wide. This makes it difficult:

- for human and applications attempting to manipulate POD content. For example, languages such as XPath 1.0 are currently not applicable to PDF documents;
- for machines attempting to learn of extraction rules automatically. In fact, existing wrapper induction approaches infer the regularity of the structure of PODs only by analyzing their internal structure.

The effectiveness of manual and automated wrapper construction is thus limited by the need to analyze the internal encoding of PODs with increasing structural complexity. The intrinsic print/visual oriented nature of PDF encoding poses many issues in defining 'ad hoc' information extraction approaches.

In the literature a number of spatial query languages for Web pages, query languages for multimedia databases and presentations, visual Web wrapping approaches, and PDF wrapping approaches, have been proposed. However, so far, these proposals provide limited capabilities for navigating and querying PODs for information extraction purposes. In particular, existing approaches are not able to generate extraction rules that are reusable when the internal structure changes, or for different documents in which information is presented by the same visual pattern. Information extraction approaches are needed that can exploit the presentation features of PODs.

ICAR-CNR is addressing these problems through the definition of spatial and semantic wrapper induction and querying approaches that allow users to query PODs by exploiting the visual patterns provided in the presentation. These approaches are grounded on document layout analysis and page segmentation algorithms combined with techniques for automatic wrapper induction and spatial languages like SXPath, a spatial extension of XPath 1.0. The innovative approaches for information extraction from PODs now being studied at ICAR-CNR permit: (i) the analysis of document layout and recognition of complex content structures like tables, sections, titles, data records, page columns, etc.; (ii) the automatic learning of extraction rules and creation of wrappers that enable relevant information to be extracted from documents such as records and objects belonging to specific classes; (iii) the navigation and querying of both Web and PDF documents by spatial primitives that exploit the spatial arrangement of content elements resulting from documents presentation.

A CNR spin-off and start-up company, Altilia srl, will implement the approaches defined at ICAR-CNR. Altilia will provide semantic content capture technologies for the content management area of the IT market.

Link: http://www.altiliagroup.com

Please contact: Massimo Ruffolo, ICAR-CNR, Italy
E-mail: ruffolo@icar.cnr.it