# Activity Report 2013

# Team DICE

# Data on the Internet at the Core of the Economy

# Table of contents

# Team DICE

**Keywords:** Web, Big Data, Socio-economic Models, Economics Of Networks, Personal Data, Electronic Voting

*Creation of the Team:* 2013 February 01.

# 1. Members

**Research Scientist**
Stephane Grumbach [Team leader, Inria, Senior Researcher, HdR]

**Faculty Member**
Stéphane Frénot [INSA Lyon, Professor, HdR]

**Engineer**
Damien Reimert [Inria]

**PhD Student**
Étienne Brodu [CIFRE Atos]

**Administrative Assistant**
Gaelle Tworkowski [Inria (20%)]

# 2. Overall Objectives

## 2.1. Overall Objectives

The Dice team has been created in february 2013 as an "action exploratoire" of Inria to initiate a multidisciplinary research on the economy of data resulting from the digital revolution and its impact on all sectors of our society including its political organisation.

With the growth of Web 2.0 systems, social data constitute a fundamental resource of the economy, much like raw materials. A resource, which is becoming as important as crude oil, and without which our societies would stop working. Data are harvested and transformed by industries that grow at an unprecedented pace. Web corporations offer extremely valuable services, which attract users in the hundreds of millions. These corporations generate ecosystems, which become as essential as public utilities and support millions of developers. The new utilities also challenge societies by making obsolete fundamental aspects of their organization, and by generating new (im)balances at global scale.

The objective of Dice is to study the complex dependencies between technological, social and economic systems of the digital age, and to propose technical contributions as well as socio-political analyses. We ambition to further investigate the impact of the digital revolution on political systems, anticipated by the French philosopher Michel Serres, as expressed in Inria's 2020 Plan.

> *"if the vast volume of global data [···] were to become accessible to as many people as possible [···], such an event would be liable to put political institutions and the sciences that study them on a new path, perhaps more quickly than we expect."*

Michel Serres also insists on the role of computer scientists to study this revolution and its social impacts.

Our contributions target both technical and theoretical aspects of the economy of personal data. Our aim is to

- study from technological as well as social, economic, political, and geopolitical points of view, the new ecosystems emerging from the services of the information society based on mediating social data, which are reshaping the very form of our organizations;
- contribute to improve the knowledge of the information society and its implications among specialists as well as non specialists, in the public opinion as well as at the political level;
- propose technological solutions that answer some of the challenges faced by our societies, such as the concentration of data, the resulting asymmetry of information, and the subsidiarity of computation, that could contribute to better distribute the knowledge among stakeholders.

## 2.2. Highlights of the Year

The team has been launched this year and has gained some visibility after a tribune in the French daily Le Monde which obtained more than 1500 "like" on the day of its publication.

*"Les données, puissance du futur", S. Grumbach, S. Frénot, Le Grand Débat, Le Monde, 8 janvier 2013*

# 3. Research Program

## 3.1. Introduction

Our aim is to address both

- challenges in the field of information technology, as well as
- transdisciplinary issues emerging from the global impact of the digital revolution.

We believe that addressing both directions at the same time is an efficient way to be relevant in each of them.

We focus on intermediation platforms, which are becoming dominant systems in the Web industries. Intermediation platforms are systems which offer services to their users, which are well tuned for their expectation, thanks to the knowledge the platform has accumulated on usage. Search engines, social networks are examples of intermediation platforms. They ensure a gatekeeping function, always in direct contact to their users, providing them with the most relevant information or contact. Their economic model relies on a biface economy, with two types of users, one subsidizing the other. Their impact goes beyond the Web, and they impact step by step all sectors of the economy, transportation, Press, education, to name a few.

So far as IT is concerned, we focus on the technologies used for intermediation, which are at the basis of the largest online systems. For the transdisciplinary questions, we focus mostly on the new equilibria that are resulting from the evolution of power balances due mostly to intermediation platforms.

## 3.2. Intermediation technologies

Dice focuses on intermediation platforms because of the central role they play in the new economy.

Intermediation platforms connect users to one another, or users to services with a very high accuracy. They rely on innovations both technological and social, which were unthinkable only ten years ago, when Facebook started. They allow communication and interaction between billions of users, gathered in the same digital space, both producers and consumers of data and services. State-of-the-art intermediation platforms include Facebook, Google, Twitter, Github, as well as Wikipedia, StackOverflow or Quora. These systems share common design and their market penetration follows the same pattern. They are built around an initial minimal viable product based on a somehow naive low-tech implementation, which evolve after a few years of improvement to Web giants. Their domination now contributes to standardize the web industry, with in particular:

- Gatekeeping, a direct relation with users together with services satisfying users' needs;
- Continuous data flows mapped to users' profiles;
- Search engines associating, in a relevant manner, producers, consumers and services.

These common characteristics lead to new software architectural standards, which are shared by all these systems, and used in the peripheric services developed in the ecosystem around their API:

- Authentication systems: openId, OAuth, ...

- Objet graphs: opengraph, follower/followee scheme, ...

- DataFlow engines: Twitter storm, Google millwheel, ...

- Databases: noSql, keyValues stores, ...

- WebBrowsers: javascript, dart, MEAN (Mongo, Express, Angular, Node),...

These architectural components impact all the digital world. Dice targets systems that use standard architecture services but preserve some aspects we consider as disruptive ones: *data concentration, data symmetry* and *computational subsidiarity*. Our current research activity includes the following directions:

- Peer-to-peer design for preserving users' primary data;

- Third parties based organic systems providing subsidiary data computation hosted at peer sites;

- In-Browser applications that impact mobile device and demonstrate instantaneous usability;

- Flow-based computing enabling a stream based exchange of information between peers at runtime.

## 3.3. Economy of the digital world

The digital revolution is impacting all sectors of our societies and organisations, education, energy, tranportation, health, to name a few. This revolution results in the phenomena of Schumpeter's *creative destruction*, with the disparition of traditional sectors and the creation of new ones. Our societies, which did not anticipate the depth of the changes, have to struggle to adapt to the pace of the development of the industry. Legal reforms in various important sectors including taxation are at stake. Some countries, more reactive than others, are clearly pulling the changes, exploiting the benefits for businesses and the capacity to generate information and value, while others are trying to catch up with the global trends.

Data form the bricks of the information society, and their flows between users and services constitute the blood of the industry. We focus in Dice on the strategic role of data in this revolution, and in particular on the systems that harvest the data and concentrate it.

We are also interested in the global political impact of this revolution, which deeply changes the relations between governments and citizens. If the privacy is the focus of considerable attention, together with the state surveillance, in Europe in particular, it is only one aspect of the new knowledge made available. Social media produce considerable knowledge not only on individuals, but on populations as well, their economic fate, their political orientation, etc. On the other hand, open data from governments allow citizens to monitor the action of their governments, as well as to contribute to it. The digital revolution, with the capacity to access information in ways unthinkable in the recent past, modifies completely the balance of powers between citizens, states and corporations.

A new trend is emerging in the humanities, around in particular the digital studies, which promote the cooperation between computer scientists and specialists of social sciences. Among them, the Berkman center for Internet and Society in Harvard, the Medialab at MIT, or the Web Science Institute in the UK have gained strong visibility. They address positive as well as negative externalities of IT for societies, that is the new potentials offered as well as their risks. The Center for Information Technology Research in the Interest of Society in Berkeley also adresses fundamental political impacts on democracy, which can be enhanced by open data as well as another philosophy of political power as currently implemented in the State of California for instance. The Open Data Institute in the UK is also a leading center for political issues in Europe. France should catch up on these research trends, at the intersection of different scientific fields.

# 4. Application Domains

## 4.1. Standard software stack for Intermediation

As linux has emerged as the reference stack for operating systems, LAMP (Linux, Apache, Mysql, Php) has emerged for the pre-Web 2.0 stack. With the emergence of intermediation system, such as those used by Facebook and other intermediation platforms, new reference stacks are used, with open architecture, which must enable the development of new intermediation businesses in a matter of days of development. Most of the tools have been developed. We are confident that the low-level toolbox is mostly designed and is based on JavaScript both at the client and the server parts, based on nosql databases such as redis or mongodb at the data layer, based on web development framework at the client and the server side, and finally on social network plugins at the intermediation layer. MEAN (MongoDB, Express, AngularJS, NodeJS) is a first proposal towards the kind of software stack we focus on, that is not exclusively devoted to intermediation purposes. We propose our own stack, based on these toolboxes to handle the future intermediation systems we envision.

## 4.2. Intermediation systems

Intermediation systems are going to govern most of our activities. Intermediation systems link all people and provide them with the best services, the most appropriate to them. They exist currently in the realm of the Web 2.0 systems such as search engine, social networks, blogging, etc. related to accessing knowledge and communicating or exchanging with people. In the near future they will make their ways in most of our systems, energy, transport, education, employment, etc. We believe that political systems will evolve as well, with a new interaction between governing bodies and citizens. The surveillance programs that are currently widely debated give an increased information on their citizens to government as well as to corporations. The trend towards open data will provide information to citizens on government actions, to an extend that we probably fail to understand today. Intermediation platforms will play a crucial role to carry on the right information or service to the right people. Our research is devoted to better understand these challenging evolutions, and propose solutions to specific aspects, in particular in the realm of elections. (cf http://www.inriality.fr/vie-citoyenne/open-data/geopolitique/va-t-delocaliser-aussi-nos/)

# 5. Software and Platforms

## 5.1. GPeer: a peer-to-peer javascript communication library

Our software development has been oriented towards systems working in browsers, with the support of an Inria ADT project in cooperation with the ASAP team. To answer our technological objectives, we are working on decentralized architectures, browser to browser, developed in javascript/HTML5. We rely on the WebRTC JavaScript protocol proposed by Google to develop a communication layer between peers. Many peer-to-peer protocols share common elements, that we group in a generic library for developing peer-to-peer systems. The joint library developed with the ASAP team handles any gossip based communication overlay. We design peer messages, tracker management and resilient behavior. The library is a standard bridge between complex browser to browser applications and low level networking layers such as WebRTC. With the use of our library, we can reproduce systems such as bitTorrent, but also provide new applications without the need of either native applications or identified servers.

## 5.2. Fluxion: a software plugin for flows in AngularJS

The joint project with Worldline aims at managing mobile code in complex Web architectures. Load variation in data-centers is currently poorly resolved. Most of the time, systems overestimate resource consumption in order to absorb burst usage. These consumption overestimation has a cost both in terms of the SLA negotiated with the client and the non-availability of reserved resources. With Wordline we focus on code mobility for high performance Web architectures and design a fast and reactive framework, transparently moving functions between running systems. The Fluxion model is our approach to design mobile application modules that are a mix of functional programming and flow based reactive systems.

## 5.3. BitBallot: a decentralized voting protocol

The BitBallot voting protocol is designed to target large scale communities. The protocol allows users to share only restricted amounts of their data and computation with central platforms as well as other peers. Convinced by the need of new election mechanisms, to support emerging forms of more continuous democracy, we are developing BitBallot, to allow elections to be organized independently of any central authority. The protocol guarantees the following properties, anonymity of the data sources, non interruptible run-time, global access to results, and non predictability of results through partial communication spying.

# 6. New Results

## 6.1. Economy of the world data flows

We have attempted to measure data flows in the world to estimate the concentration of the data industry. It is well known that the main plateforms of the Web, Google, Facebook, Amazon, etc. are concentrated in a few countries, mostly in the USA. Some countries, mostly asian, such as China, Russia, Korea or Japan have succesfully developed their own Web 2.0 industry, while others, such as European countries, hava failed to do so. We have explored in [6], the strategy of China, which has the largest Web industry behind the US and has made a priority of keeping its data at home, with systems in all activity sectors developed in general only one or two years after their main americain counterparts. The innovation strategy of China aims in all fields to achieve technological independence, with at most 30% of foreign IP as we have shown in [2].

A tentative measure of the flows of personal data from different regions is proposed in [3], based on the traffic on the largest platform at the international level. We show in particular that personal data captured in Europe on Web platforms mostly go to the US industry. In [4], we investigate the invisible part of the Web constituted by the trackers that are hidden on Web pages and transfer data to third parties, and show that the domination of the US is even stronger on trackers than it is on the visible Web.

## 6.2. Flow systems

We are currently working with Bull SA, Manuel Selva (PhD) and Lionel Morel from the Socrates team to build a monitoring framework for dynamic data-flow system in many-core architectures. Data-flow computing models computation as a pipeline of computation units absorbing a continuous stream of data. This computing model suits application development for embedded devices such as MPEG-4 video encoders. The incoming data flow is sliced into small size token (e.g. video frames). Each time, all computational units take some tokens from their inputs and produce some tokens on their outputs. We focus [7], [8] on a management layer for handling dynamic dataflow programs in many-core architectures, where computation units may be relocated at runtime from one core to another. The questions raised by Twitter Storm, Google Millwheel or Yahoo S4, are in essence very similar. Can our current architectures hold the information dataflow produced by users in terms of computing power and memory usage? We are currently extending these embedded results to study dataflow architectures with ATOS on flow computing inside Web browsers.

François Goichon will defend his PhD on resource access equity into best-effort operating systems such as Linux. Linux is built over a layered architecture, where each layer owns a local policy that may lead to a global policy being far from best-effort. With Guillaume Salagnac from Socrates team, we show [5], [9] that we can develop malware user space applications exploiting embedded linux firmware and device drivers differential policy that can block other concurrent applications from accessing CPU time. When this kind of applications are installed in multi-tenant architectures as found in cloud shared space, it can slowdown the entire system. These results are interesting for Dice when considering access time in web browser. Current in-browser applications are developed in Javascript, which imposes a single threaded executed model to the developer, yet operated on a multi-core architecture. Best-effort operating systems are not the best approaches to handle flow based applications that become the norm, and we think that some small, low-level shifts, should be considered.

# 7. Bilateral Contracts and Grants with Industry

## 7.1. Bilateral Contracts with Industry

Dice has bilateral contracts with three companies.

**Worldline** Wordline is a leader in B2B applications development, and is in the front line to provide new technical solution in the Web 2.0 era. We have a CIFRE partnership contract on the study of flow based architectures both at the data centers and at the Web browser level.

**OrangeLabs** We are finishing a joint work with OrangeLab on smart buildings and the management of home equipments. The project aims to provide a digital management layer for long living equipment that do not have network connexion for historical or technical reasons. The collaboration relies on a CIFRE partnership.

**BullSA** BullSA is producing and designing next generation Many-Core architecture. Although most of the time these calculators are used in real-time, closed environment such as military equipments, the dynamic, adaptability, and upgradable nature of systems is a real issue. We participate in a joint project to design a management layer for handling dynamic data flow application in a soft real-time context.

# 8. Partnerships and Cooperations

## 8.1. Regional Initiatives

Dice is involved in a regional project of the Rhône-Alpes region, ARC6 "Innovative Services for Social Networks", with Telecom Saint Etienne.

## 8.2. National Initiatives

### 8.2.1. ANR

Dice is involved in two new ANR projects, to start at the end of 2013,
- C3PO, on Collaborative Creation of Contents and Publishing using Opportunistic networks, with LT2C Telecom Saint-Etienne, INSA LYON, IRISA, ChronoCourse, et Ecole des Mines de Nantes.
- Socioplug, Social Cloud over Plug Networks, Enabling Symmetric Access to Data and Preserving Privacy, with LINA / Université de Nantes, Université de Rennes 1, INSA Lyon.

## 8.3. European Initiatives

### 8.3.1. FP7 Projects

Dice is involved in the CSA project "Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE)", Objective ICT-2013.4.2 Scalable data analytics (c) Societal externalities of Big Data roadmap.

# 9. Dissemination

## 9.1. Scientific Animation

The digital revolution induces rapid changes in our societies, that are often not well prepared to adapt and take full advantage of them. Our objective is to communicate broadly to non specialist communities, through teaching, conferences in other communities, as well as through the media.

S Frénot and/or S Grumbach have been invited speaker or panelist at the following conferences in IT related fields:

- Panel La politique des données personnelles : big data ou contrôle individuel ?, IXXI, ENS de Lyon et Villa Gillet (Festival Mode d'Emploi), Lyon, 21 novembre 2013.
- Chair and organization of Session Managing the flow computing, Fossa, Lille, novembre 2013. Slides and presentation for this session are available at https://fossa.inria.fr/slides/
- Tutoriel Les flux de données personnelles, enjeux technologiques, économiques et stratégiques, 29e Journées Bases de Données Avancées, Nantes, 22-25 Octobre 2013.
- Panel: Big Data : c'est aussi un sujet de sécurité, Les assises de la sécurité et des systèmes d'information, Monaco, 2-5 octobre 2013.
- Invited speaker Congrès Big Data, Turning the Data Deluge into Decisions, CNIT, Paris, avril 2013
- La France à la périphérie de la société de l'information? Café techno, Paris, mars 2013
- Panel on Big Data, 6th International Conference on Computers, Privacy and Data Protection, CPDP, Reloading Data Protection, Brussels, january 2013

## 9.2. Teaching - Supervision - Juries

### Teaching

We have been involved in the following courses:

- INSA, Frénot, Grumbach, TweetMyFace: A 38 hours optional course on social network technical architectures, evolutions and designs. The course covers current Web technologies from low-level networking to Facebook and Twitter API. L2 (Since 2012)
- INSA, Frénot, Agility: A 32 hours optional course on agile software development presenting both iterative (SCRUM) and stream based approaches (LEAN IT). M1. (since 2011)
- INSA, Frénot, Innovating Project: Supervising 250 hours student project aim at managing innovating projects. Each student group leads its own subject during one semester. All projects and organizational details are publicly available here: http://tc-pi.insa-lyon.fr M1 (Since 2006)
- INSA, Frénot, Learn Other Languages: The aim of the course is to improve one's skills in current state-of-the-art programing and discover different ways to develop using mainly web-oriented programming languages. M1 (new course)
- INSA, Frénot, Innovation and Transfer for Software: This final year optional course targets a specific activity for engineer profession dedicated to software transfer from research labs to industrials.

Stéphane Grumbach supervised the following thesis:

PhD: Ahmad Ahmad-Kassem, Programming Networks with Intentional Destination, Université de Lyon, november 2013

Stéphane Frénot supervised the following thesis:

- Etienne Brodu started in 2013. Worldline partnership CIFRE Phd. "Flow-base operating systems and programing languages. A new deal for social network architectures"
- François Goichon defense december 2013. French Ministry of research. "Resource access equity for best-effort shared operating systems"
- Zheng Hu defense january 2014. Orange-Labs partnership CIFRE PhD. "Self-configuration, monitoring and control of physical entities via Sensor and Actuator Networks"
- Dan Yufang defense Mid 2014. CSC China. "Secure and healable usage of components in a dynamic service-oriented architecture-based system"
- Manuel Selva defense End 2014. Bull SA partnership. "Monitoring data-flow programs"

**Juries**

Stéphane Frénot was involved in the following committee as reporting member.

- Roberto Minerva: Will the telco survive to an ever changing world? Technical considerations leading to disruptive scenarios.
- Aurelien Faravelon: A privacy aware conceptual and implementation framework for service oriented architecture based on access control.
- Azzedine Amiar : Trace and log analysis in micro-controllers.

## 9.3. Popularization

Intervention in other arenas:

- Co-organization of the CARA community, that gather IT professionals around agility design. Meetings occurs once per month from 7PM to 10PM and gather around 50 people to discuss and debate. The complete animation scheme is available at http://lyon.clubagilerhonealpes.org/
- La révolution numérique, L'enseignement philosophique et les sciences: nouvelles perspectives, Fondation Simone et Cino del Duca, Paris, 13 novembre 2013
- Panel: Open Data and Civic Participation, Can "Open Data" Improve Democratic Governance?, CITRIS Data & Democracy Initiative and Institute of Governmental Studies, Berkeley, 12 September 2013.
- Big Data: Where Does Europe Stand? a Citizen's Controversy, at the Madariaga – College of Europe Foundation, with Roberto Viola, Deputy Director of DG Communication, Brussels, 12 July 2013.
- La Révolution numérique, un enjeu politique, Conférence au Cercle Pierre Mendes France de Lyon, 25 juin 2013
- Les données, nouveau moteur de l'économie, Congrès France Génétique Elevage, janvier 2013

Intervention in political instances. The Dice team aims at interacting with political representatives at the French Assemblée nationale, the Sénat, as well as instances of the European Union.

- Global perspective on the information society, Invited speaker at the Council of the European Union, Part I Europe at the periphery of the information society? Brussels, April 17, 2013, Part II Information society in China, the Beijing consensus? Brussels, May 14, 2013,
- La dépendance de la France en matière de données et services numériques, Assemblée Nationale, Audition publique du 21 février 2013, "Le risque numérique : en prendre conscience pour mieux le maîtriser ?",

# 10. Bibliography

## Publications of the year

### Doctoral Dissertations and Habilitation Theses

[1] A. AHMAD KASSEM. , *Programming Networks with Intensional Destinations*, INSA de Lyon, November 2013, http://hal.inria.fr/tel-00908354

### Articles in National Peer-Reviewed Journals

[2] G. FABRE, S. GRUMBACH. *Le "vrai Grand Bond en avant" La politique de recherche et d'innovation en Chine*, in "Le Débat", February 2013, http://hal.inria.fr/hal-00911187

[3] S. FRÉNOT, S. GRUMBACH. *Les données, objets de toutes les convoitises*, in "Hérodote", February 2014, http://hal.inria.fr/hal-00908894

### International Conferences with Proceedings

[4] C. CASTELLUCCIA, S. GRUMBACH, L. OLEJNIK. *Data Harvesting 2.0: from the Visible to the Invisible Web*, in "The Twelfth Workshop on the Economics of Information Security", Washington, DC, United States, Allan Friedman, June 2013, http://hal.inria.fr/hal-00832784

[5] F. GOICHON, G. SALAGNAC, S. FRÉNOT. *Swap Fairness for Thrashing Mitigation*, in "ECSA - European Conference on Software Architecture - 2013", Montpellier, France, July 2013, 17 p. , http://hal.inria.fr/hal-00912875

[6] S. GRUMBACH. *The stakes of Big Data in the IT industry: China as the next global challenger?*, in "The 18th International Euro-Asia Research Conference, The Globalisation of Asian Markets: implications for Multinational Investors", Venise, Italy, January 2013, http://hal.inria.fr/hal-00745758

[7] M. SELVA, L. MOREL, K. MARQUET, S. FRÉNOT. *Extending dataflow programs for guaranteed throughput*, in "MES '13 - International Workshop on Many-core Embedded Systems", Tel Aviv, Israel, ACM, June 2013, pp. 54-57 [*DOI :* 10.1145/2489068.2489077], http://hal.inria.fr/hal-00876566

### National Conferences with Proceedings

[8] M. SELVA, L. MOREL, K. MARQUET, S. FRÉNOT. *A QoS Monitoring System for Dataflow Programs*, in "COMPAS 2013 - Conférence d'informatique en Parallélisme, Architecture et Système", Grenoble, France, N/A, January 2013, http://hal.inria.fr/hal-00780976

### Research Reports

[9] F. GOICHON, G. SALAGNAC, S. FRÉNOT. , *Swap Fairness for Thrashing Mitigation*, Inria, July 2013, n$^{\text{o}}$ RR-8388, 17 p. , http://hal.inria.fr/hal-00879084