UNIVERSITY OF Southampton

WSRI

web science research initiative

Contents

of this presentation

- introduction to WSRI
- what we do
- what is Web Science
- exemplars
- current activities

research / thought leadership / education

WSRI
web science research initiative

Introduction

our motivation

- the Web has been transformational
- we need to understand it
- anticipate future developments
- identify opportunities and threats
- we have established a new discipline: Web Science

research / thought leadership / education

WSRI
web science research initiative

Introduction

our aims

- promote and encourage multidisciplinary collaborative research to study the development of the Web
- provide a global forum to enable academia, government and industry to understand the scientific, technical and social factors that drive the growth of the Web and enable innovation
- devise curricula for the new discipline of Web Science so as to train future generations of Web Scientists

research / thought leadership / education

WSRI
web science research initiative

Introduction

directors

Tim Berners-Lee

Wendy Hall

Nigel Shadbolt

Daniel Weitzner

the reputations, experience and skills of our Directors enables us to work closely alongside academia, government, industry and donors to realize our aims

research / thought leadership / education

WSRI
web science research initiative

Where we are

launch November 06

"Web Science represents a pretty big next step in the evolution of information. This kind of research is likely to have a lot of influence on the next generation of researchers, scientists and, most importantly, the next generation of entrepreneurs who will build new companies from this."

Dr Eric Schmidt, CEO, Google Inc.

"Web Science research is a prerequisite to designing and building the kinds of complex, human-oriented systems that we are after in services science."

Irving Wladawsky-Berger, VP, Technical Strategy and Innovation, IBM Corporation.

research / thought leadership / education

WSRI
web science research initiative

Governance

structure

- scientific council
  leaders in their disciplines
- strategic advisory board
  individuals with influence
- corporate advisory board
  companies committed to Web Science

research / thought leadership / education

WSRI
web science research initiative

Where we are

WSRI Operational Phases

Phase 1: Nov 06 to Nov 07

- Launched the concept and targeted activities

Phase 2: Dec 07 to Nov 08

- Establish operational base in America and Europe

Phase 3: Dec 08 onwards

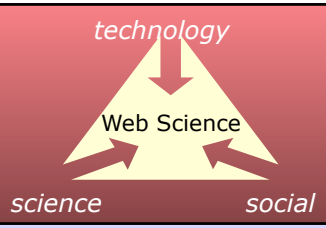
- Build a global, multi-partner organization and expand activity base

research / thought leadership / education

WSRI
web science research initiative

WSRI- a catalyst for the future web

vision



To create a multi-disciplinary approach to understand how the Web will evolve, anticipate issues that will arise and proactively address these issues to ensure that the Web remains as a positive enabler for our society

how

Research

- Evolution of the web
  - Structure
  - Protocols
  - Social effects
  - Fragmentation
  - Rates of change
- Applications
  - Education
  - Science
  - Health
  - eGovernment

Thought leadership

- Global forum of experts to inform
  - Academia, Government, Industry
  - Raise awareness, lead thinking, disseminate
- Workshops
  - Issue generation
  - Consultancy
  - Conferences

Education

- Web Science curriculum development at undergraduate and graduate levels
- Creation of Web Science capabilities and talent
- Role of web in innovation
- Web as driver for development

- Integrated, multi-disciplinary perspective
- Capacity for new field of Web Science
- Solutions to anticipated issues
- Coordination on a global scale

impact

→ Business

- Enhanced ways of working
- Address future talent needs
- Opportunities for innovation

→ Communities

- Ability to share information
- Informed resource allocation
- Capacity building for Social welfare

→ Policy Makers

- More informed policy choices
- Broader reach of input
- More access

→ Foundations/ non-profits

- Connect resources with needs
- Facilitate social goals
- Maximize impact

What is Web Science?

The Web has been transformational

- Largest human information construct

How are we to

- Understand what it is
- Develop its engineering
- Ensure its social benefit
- Develop capacity

This requires a new interdisciplinary field -

- This field we call Web Science

research / thought leadership / education

WSRI
web science research initiative

Web Science EMERGES

Studying the Web will reveal better ways to exploit information, prevent identity theft, revolutionize industry and manage our ever growing online lives

By Nigel Shadbolt and Tim Berners-Lee

KEY CONCEPTS

The relentless rise in Web pages and links is creating emergent properties, from social networking to virtual identity theft, that are transforming society.

A new discipline, Web science, aims to discover how Web traits arise and how they can be harnessed or held in check to benefit society.

Important advances are beginning to be made; more work can solve major issues such as securing privacy and conveying trust.

—*The Editors*

Since the World Wide Web blossomed in the mid-1990s, it has exploded to more than 15 billion pages that touch almost all aspects of modern life. Today more and more people's jobs depend on the Web. Media, banking and health care are being revolutionized by it. And governments are even considering how to run their countries with it. Little appreciated, however, is the fact that the Web is more than the sum of its pages. Vast emergent properties have arisen that are transforming society. E-mail led to instant messaging, which has led to social networks such as Facebook. The transfer of documents led to file-sharing sites such as Napster, which have led to user-generated portals such as YouTube. And tagging content with labels is creating online communities that share everything from concert news to parenting tips.

But few investigators are studying how such emergent properties have actually blossomed, how we might harness them, what new phenomena may be coming or what any of this might mean for humankind. A new branch of science—Web science—aims to address such issues. The timing fits history: computers were built first, and computer science followed, which subsequently improved computing significantly. Web science was launched as a formal discipline in November 2006, when the two of us and our colleagues at the Massachusetts Institute of Technology and the University of Southampton in England announced the beginning of a Web Science Research Initiative. Leading researchers from 16 of the world's top universities have since expanded on that effort.

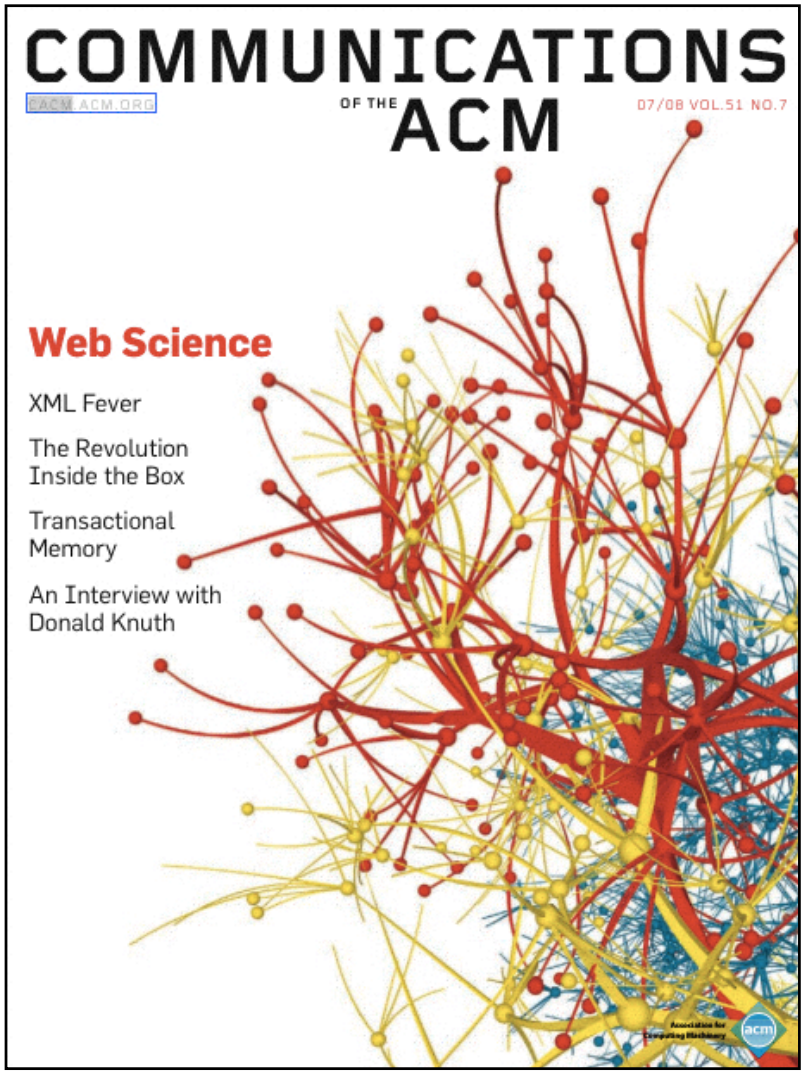
This new discipline will model the Web's structure, articulate the architectural principles that have fueled its phenomenal growth, and discover how online human interactions are driven by and can change social conventions. It will elucidate the principles that can ensure that the network continues to grow productively and settle complex issues such as privacy protection and intellectual-property rights. To achieve these ends, Web science will draw on mathematics, physics, computer science, psychology, ecology, sociology, law, political science, economics, and more.

Of course, we cannot predict what this nascent endeavor might reveal. Yet Web science has already generated crucial insights, some presented here. Ultimately, the pursuit aims to answer fundamental questions: What evolutionary patterns have driven the Web's growth? Could they burn out? How do tipping points arise, and can that be altered?

Insights Already

Although Web science as a discipline is new, earlier research has revealed the potential value of such work. As the 1990s progressed, searching for information by looking for key words among the mounting number of pages was returning more and more irrelevant content. The founders of Google, Larry Page and Sergey Brin, realized they needed to prioritize the results.

Their big insight was that the importance of a page—how relevant it is—was best understood in terms of the number and importance of the pages linking to it. The difficulty was that part of this definition is recursive: the importance of a page is determined by the importance of the



research / thought leadership / education

WSRI
web science research initiative

Contents

Practical Examples of Web Science

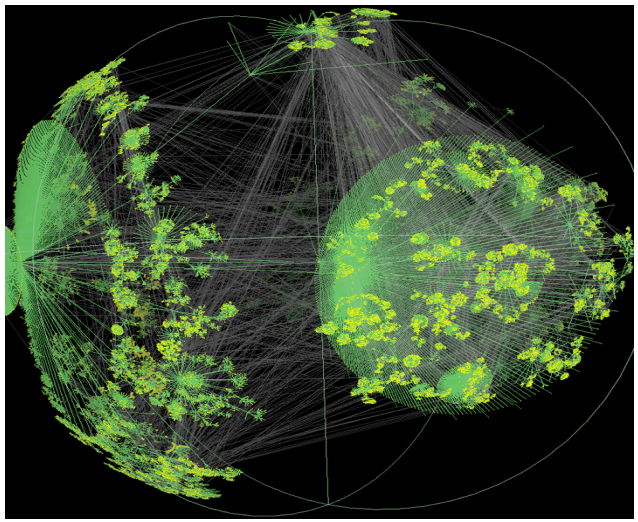
- Past
- Present
- Future

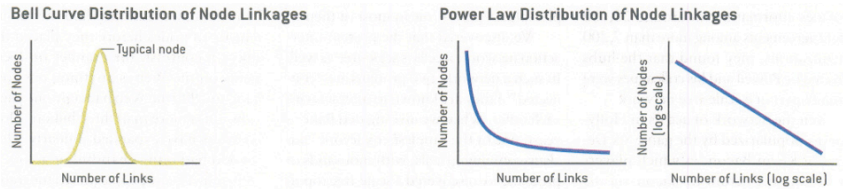
research / thought leadership / education

WSRI
web science research initiative

Web Science - Past

Web Structure





Scale-free

Some nodes are of high degree most are low degree

Structure and dynamics independent of the network size

Power laws

The degree distribution follows a power law, with an exponent $\beta > 2$.

Small worlds

The average distance (or diameter) is much smaller than the order of the graph.

Hubs and authorities

The number of distinct bipartite cliques or cores is large when compared to a random graph with the same number of nodes and edges.

research / thought leadership / education

WSRI
web science research initiative

Dynamics of the Web

The Web is different from most hitherto-studied systems in that it is changing at a rate which is of the same order as, or maybe greater than, our ability to observe it.

How are we to instrument the Web and how can we log it or identify behaviours?

Mathematical tools help to analyse the changing structure of the Web (e.g., using graph theory). The graph beneath the graph

Sociology can develop an understanding of the two-way process by which individuals and technologies shape each other.

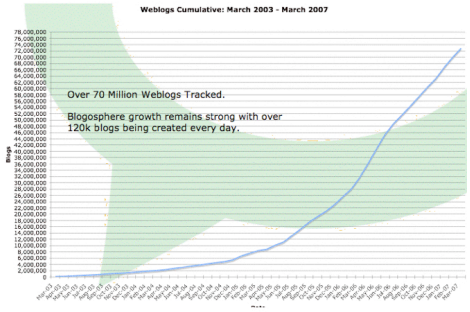
whether law is a catalyst for Web dynamics, or merely reactive to it.

language (and e.g., the preponderance of people for whom English is a second language) is affecting the development of the Web.

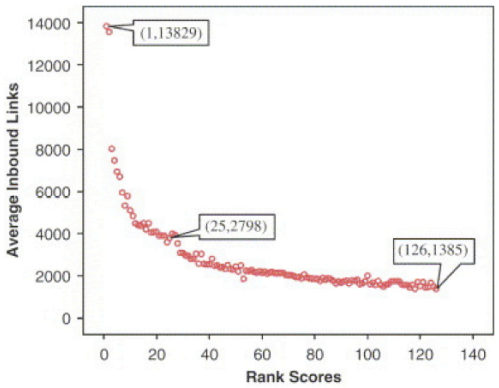
Web Science - Past

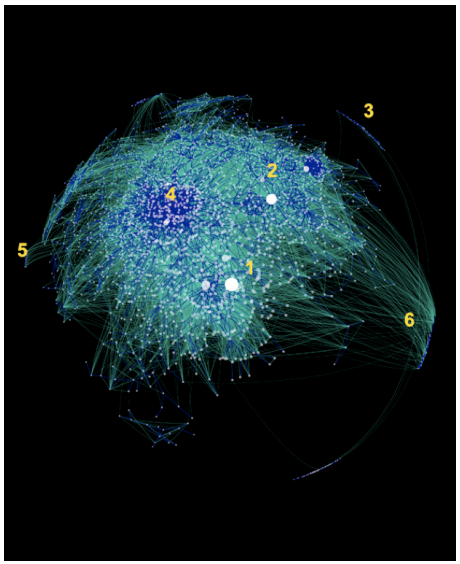
The Blogosphere

- by understanding the scientific, technical and social factors that drive the growth of the Web we can understand past Web phenomena



Top Blog sites





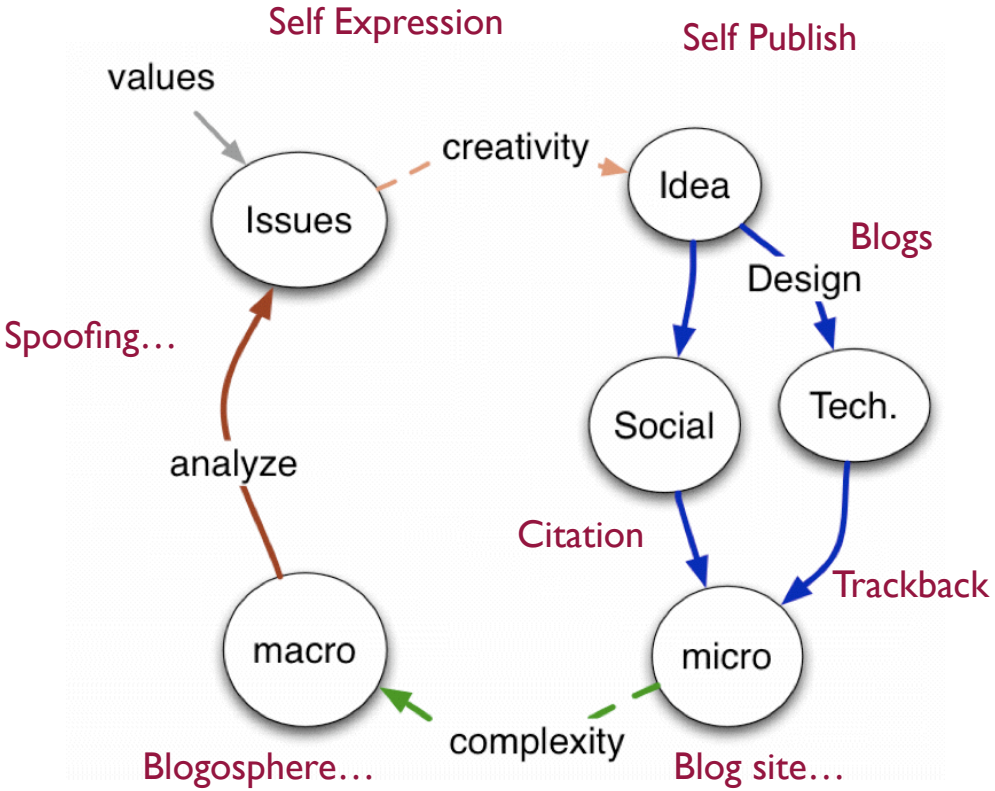
research / thought leadership / education

WSRI
web science research initiative

Web Science

Understanding

- creative innovation
- design and engineering
- the social and the technical
- interpretation and analysis



research / thought leadership / education

WSRI
web science research initiative

“Openness” on the Web

The Web, as it exists today, is a complex mixture of open, public areas and closed, private zones.

The Web must be based on open platforms v property rights provide the strongest incentive for innovation in the Web.

What is meant by “openness”?

How can legal frameworks be constructed to deal with openness on the Web?

Is openness necessary for innovation, or are private and commercial incentives more effective?

Is openness compatible with the security requirements of e.g., e-health applications?

When is it important to release intellectual property to build a user base, and when should a more restrictive business model come into place?

Web Science - Present

Wikipedia - Collective Intelligence

- Shape and structure
- Scale free
- Preferential attachment
- Communities
- Values and obligations
- Incentives

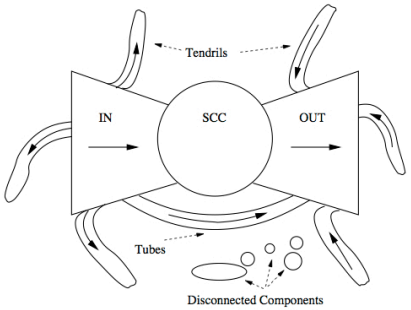
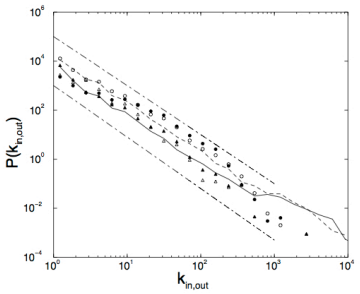
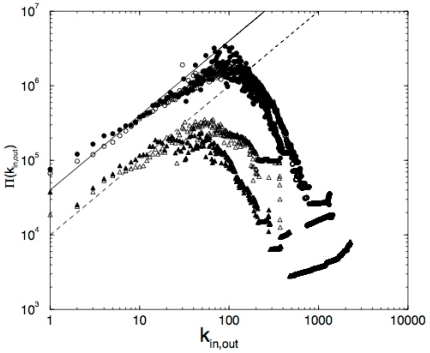


FIG. 1: The shape of the Wikipedia network





research / thought leadership / education

WSRI
web science research initiative

Web Science - Present

Wikipedia - Collective Intelligence

- Shape and structure
- Scale free
- Preferential attachment
- Communities
- Values and obligations
- Incentives

Motivation	Mean
Fun	**6.10** (1.15) [0.322***]
Ideology	**5.59** (1.71) [0.110]
Values	**3.96** (1.55) [0.175*]
Understanding	**3.92** (1.48) [0.296***]
Enhancement	**2.97** (1.39) [0.313***]
Protective	**1.97** (1.05) [0.306***]
Career	**1.67** (0.94) [0.185*]
Social	**1.51** (0.92) [0.027]

*significant at 0.05 level
**significant at 0.01 level
***significant at 0.001 level

Motivation	Question example
Protective	"By writing/editing in Wikipedia I feel less lonely."
Values	"I feel it is important to help others."
Career	"I can make new contacts that might help my business or career.
Social	"People I'm close to want me to write/edit in Wikipedia."
Understanding	"Writing/editing in Wikipedia allows me to gain a new perspective on things."
Enhancement	"Writing/editing in Wikipedia makes me feel needed."
Fun	"Writing/editing in Wikipedia is fun."
Ideology	"I think information should be free."

research / thought leadership / education

WSRI
web science research initiative

Collective Intelligence

Collaborative endeavour with only light rules of co-ordination that lead to the emergence of large-scale, coherent resources (such as Wikipedia).

How, from a technical point of view, can collective intelligence be enabled?

What are the socio-economic reasons why individuals participate in collective endeavour?

What legal framework governs (or should govern) the resources that are created?

What is the psychology of identification with an online collective community?

How can collective intelligence emerge, given the different languages used by different genders, races, classes and communities?

What role is there for policy-makers to engage in and facilitate collaborative endeavour?

Harnessing the potential of collective and collaborative intelligence is an important theme for governments as they try to engage citizens, verify their legitimacy and find creative policy levers.

the social web

- social networking sites
- (inter-linked) blogs + comments + aggregators
- community-edited news sites, participatory journalism
- content-sharing sites
- discussion forums, newsgroups
- wikis, Wikipedia
- services that allow sharing of bookmarks/favorites
- ...and **mashups** of the above services

"democratic", participatory, conversational

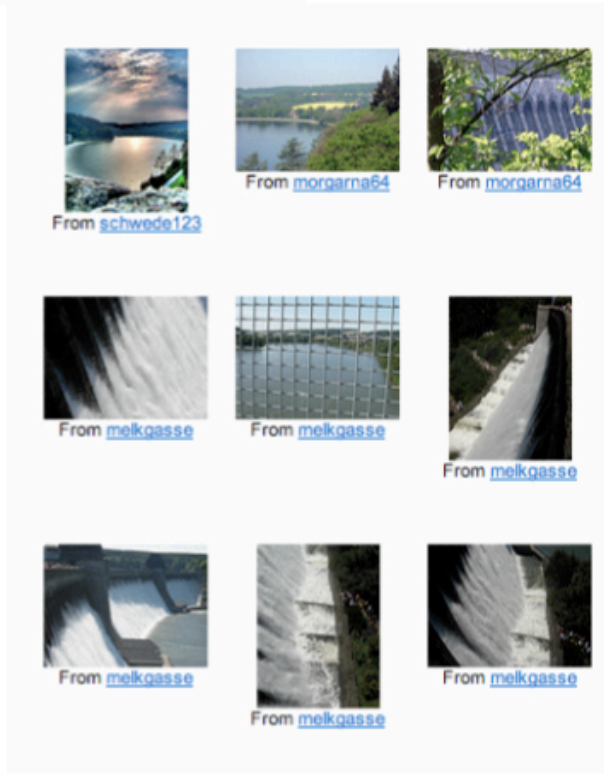
Material from Ciro Cattuto

http://isiosf.isi.it/~cattuto

research / thought leadership / education

WSRI
web science research initiative

http://flickr.com



"folksonomy"

research / thought leadership / education

WSRI
web science research initiative

Flickr's "ecology"

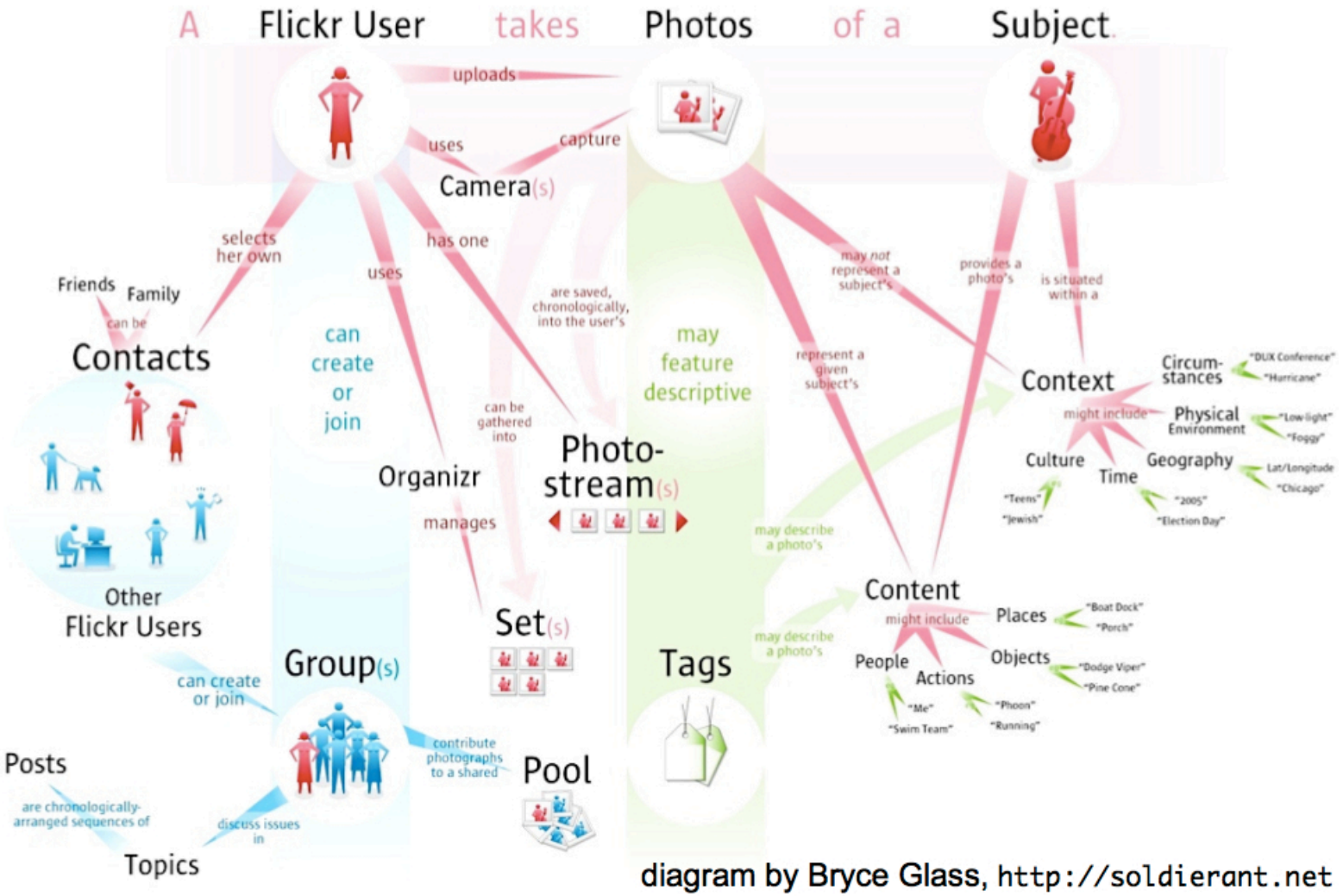


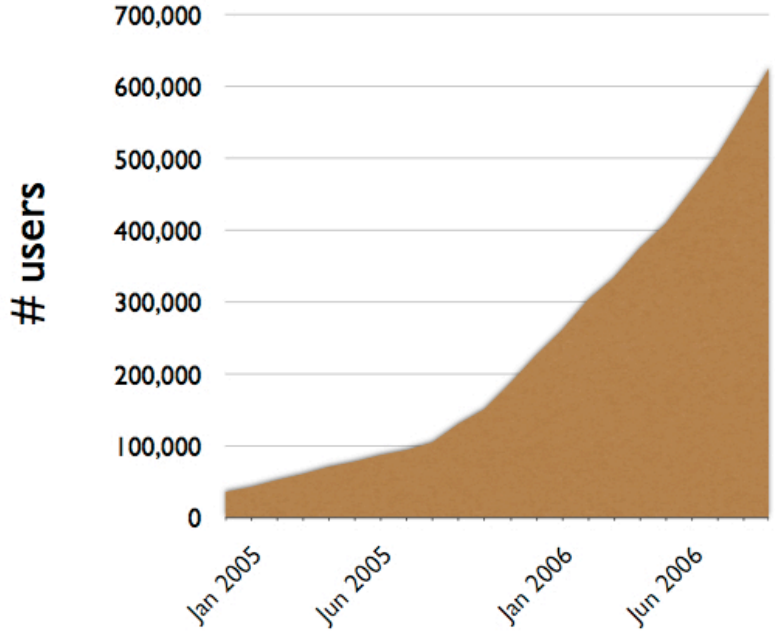
diagram by Bryce Glass, http://soldierant.net

research / thought leadership / education

WSRI
web science research initiative

experimental data

- large-scale *del.icio.us* snapshot from distributed crawl
- from beginning of 2004 up to november 2006
- full hypergraph structure + timestamps



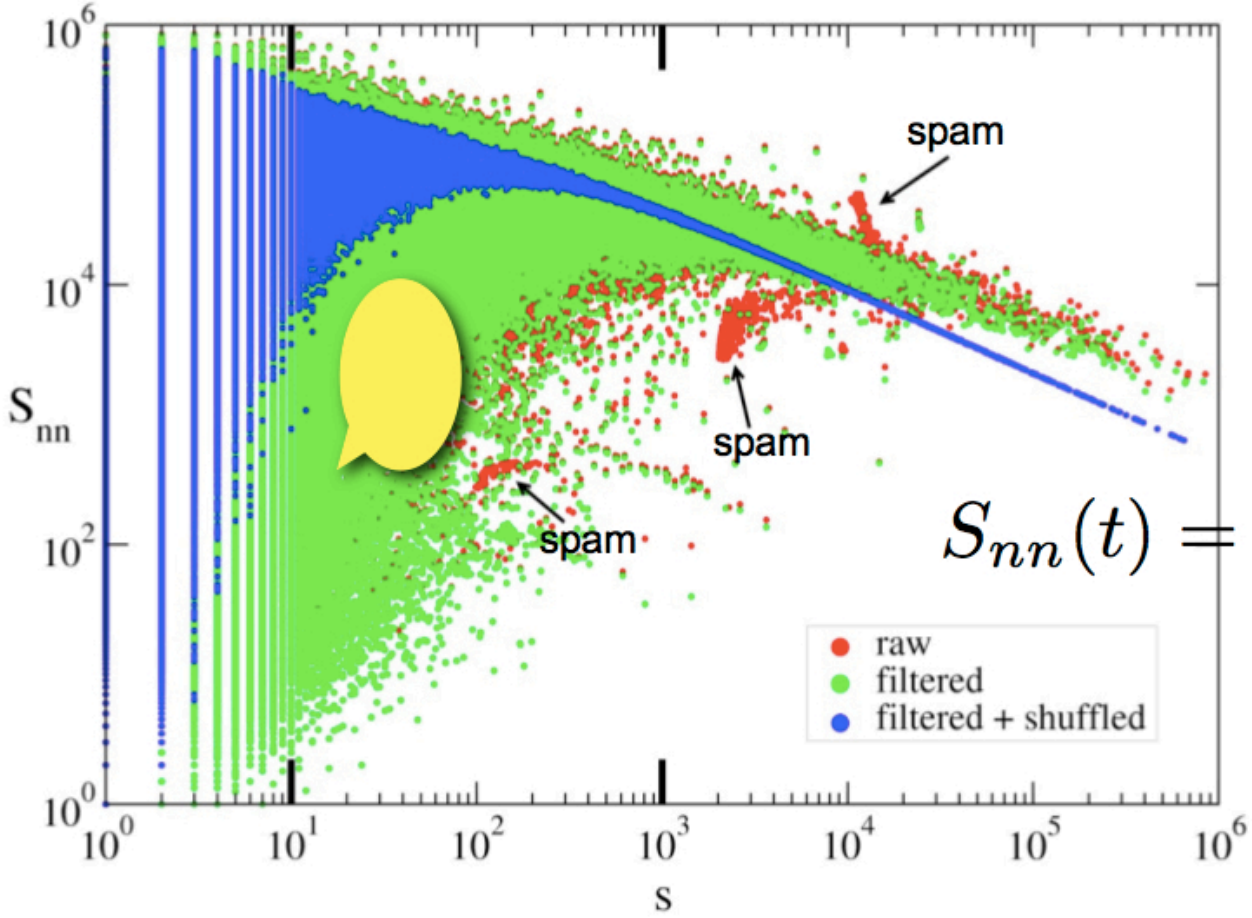
$\sim 650{,}000$ users

$\sim 2 \cdot 10^7$ resources

$\sim 5 \cdot 10^7$ posts

$\sim 3 \cdot 10^6$ *distinct* tags

networks of tag co-occurrence



$$S_{nn}(t) = \frac{1}{k_i} \sum_{j=1}^{k_i} s_t$$

Material from Ciro Cattuto

http://isiosf.isi.it/~cattuto

CC *et al.*, AI Communications **20**, 245 (2007)

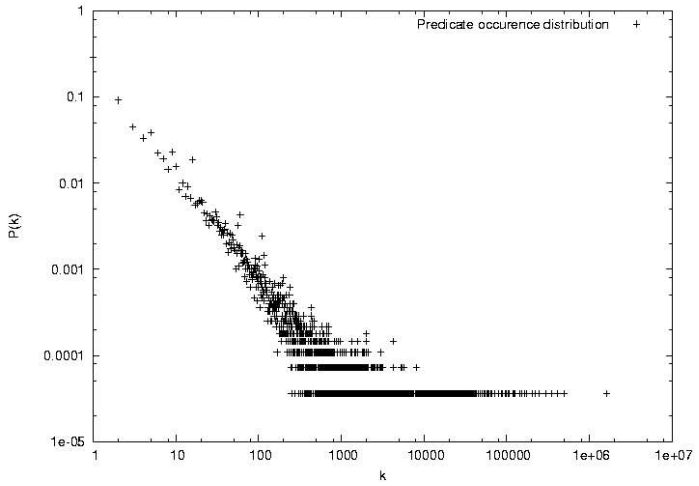
research / thought leadership / education

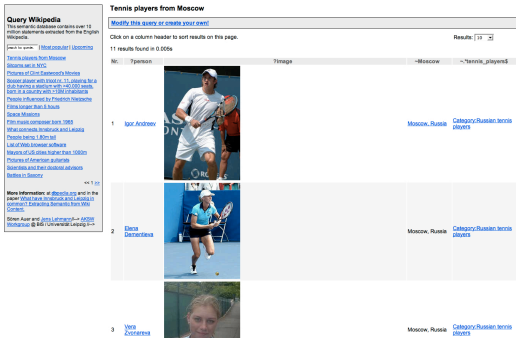
WSRI
web science research initiative

Web Science - Future

Linked data

- Is the Semantic Web at a tipping point
- It is not services it is content
- The Linked Web of Data
- See DBpedia etc.





research / thought leadership / education

WSRI
web science research initiative

“Inference”

The amount of information on the Web is enormous and growing exponentially.

It is a major challenge to measure the amount of information contained in the Web.

How we are to browse, explore and query the Web at this scale?

How inference can be supported at the Web scale;

How can context be represented and supported?

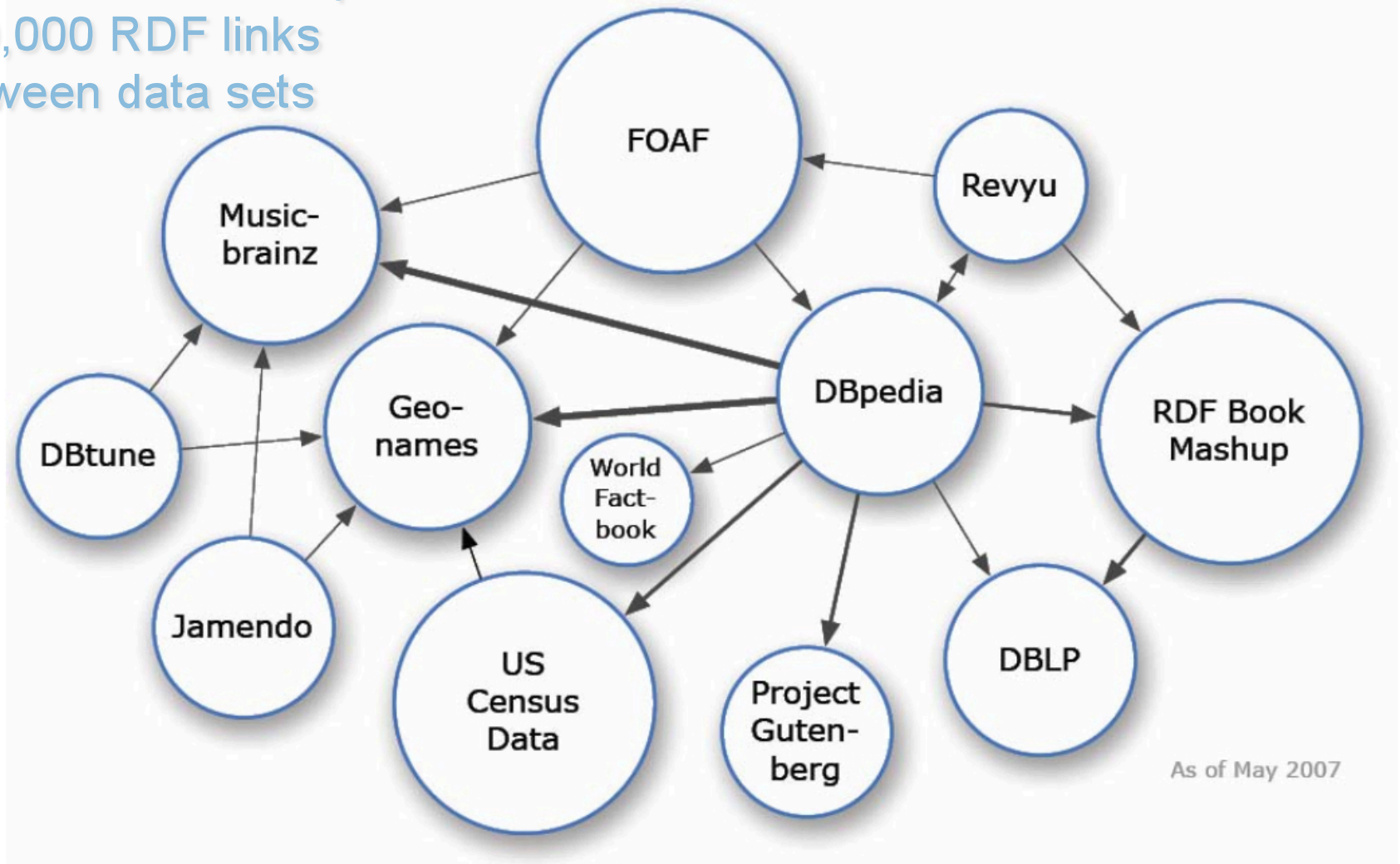
The design of interfaces for querying complex data?

How can the data sources within the Web be exploited to help us to develop understanding of the sociological aspects of the Web?

Linked Data on the Web: May 2007

UNIVERSITY OF Southampton
School of Electronics and Computer Science

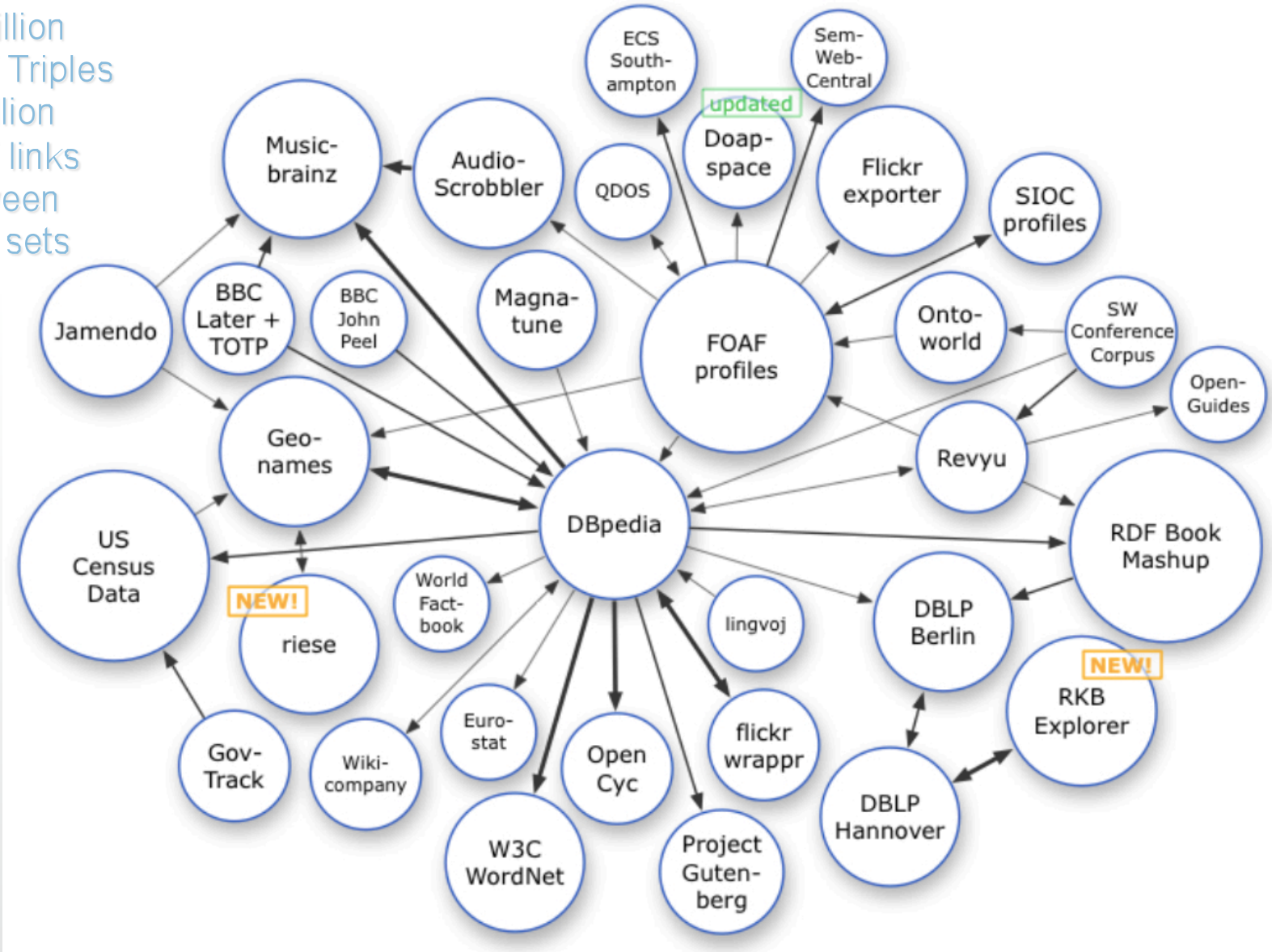
500 Million RDF Triples
120,000 RDF links between data sets



Linked Data on the Web: April 2008

UNIVERSITY OF Southampton
School of Electronics and Computer Science

23 billion RDF Triples
3 million RDF links between data sets



Content, Emergence and Unanticipated Reuse

UNIVERSITY OF Southampton
School of Electronics and Computer Science

The four micro principles of the Semantic Web

1. All entities of interest, such as information resources, real-world objects, and vocabulary terms should be identified by URI references.
2. URI references should be dereferenceable, meaning that an application can look up a URI over the HTTP protocol and retrieve RDF data about the identified resource.
3. Data should be provided using the RDF/XML syntax.
4. Data should be interlinked with other data.

Nigel Shadbolt is a professor of artificial intelligence in the School of Electronics and Computer Science at Southampton University. Contact him at nrs@ecs.soton.ac.uk.

Tim Berners-Lee is the director of the World Wide Web Consortium, a senior researcher at the Massachusetts Institute of Technology's Computer Science and Artificial Intelligence Laboratory, and a professor of computer science in the Department of Electronics and Computer Science at Southampton University. Contact him at timbl@w3.org.

Wendy Hall is a professor of computer science in the School of Electronics and Computer Science at Southampton University. Contact her at wh@ecs.soton.ac.uk.

The Semantic Web Revisited

Nigel Shadbolt and Wendy Hall, *University of Southampton*
Tim Berners-Lee, *Massachusetts Institute of Technology*

Web Science - Future

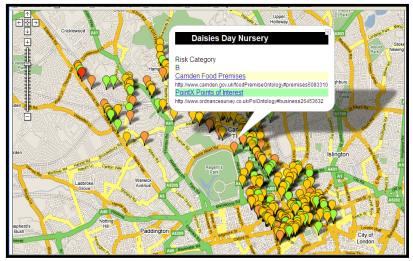
Linked data

- by understanding the scientific, technical and social factors that drive the growth of the Web we can anticipate future Web phenomena
- e.g the linked data Web

The four micro principles of the Semantic Web

1. All entities of interest, such as information resources, real-world objects, and vocabulary terms should be identified by URI references.
2. URI references should be dereferenceable, meaning that an application can look up a URI over the HTTP protocol and retrieve RDF data about the identified resource.
3. Data should be provided using the RDF/XML syntax
4. Data should be interlinked with other data.

Figure 8 Browsing the Structured Data Web for Proteomics



research / thought leadership / education

WSRI
web science research initiative

Web Science

Social impact of linked data

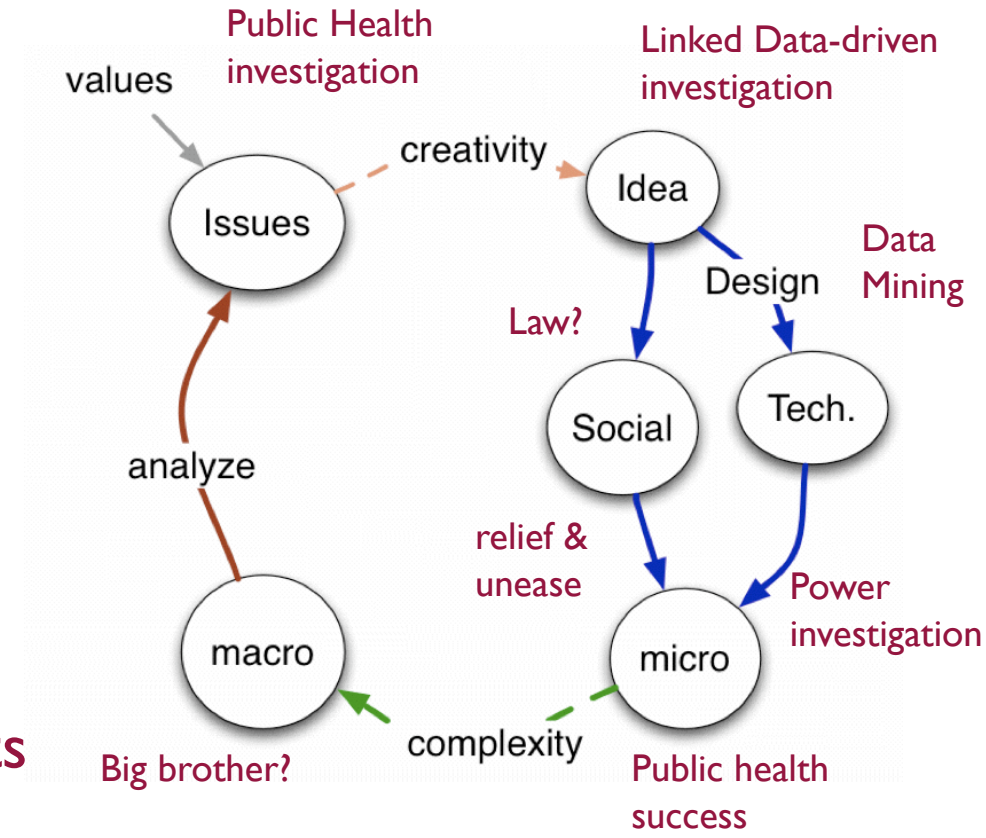
Public health challenges

Data mining power

Privacy concerns

Unclear legal rules

Unanticipated social effects



research / thought leadership / education

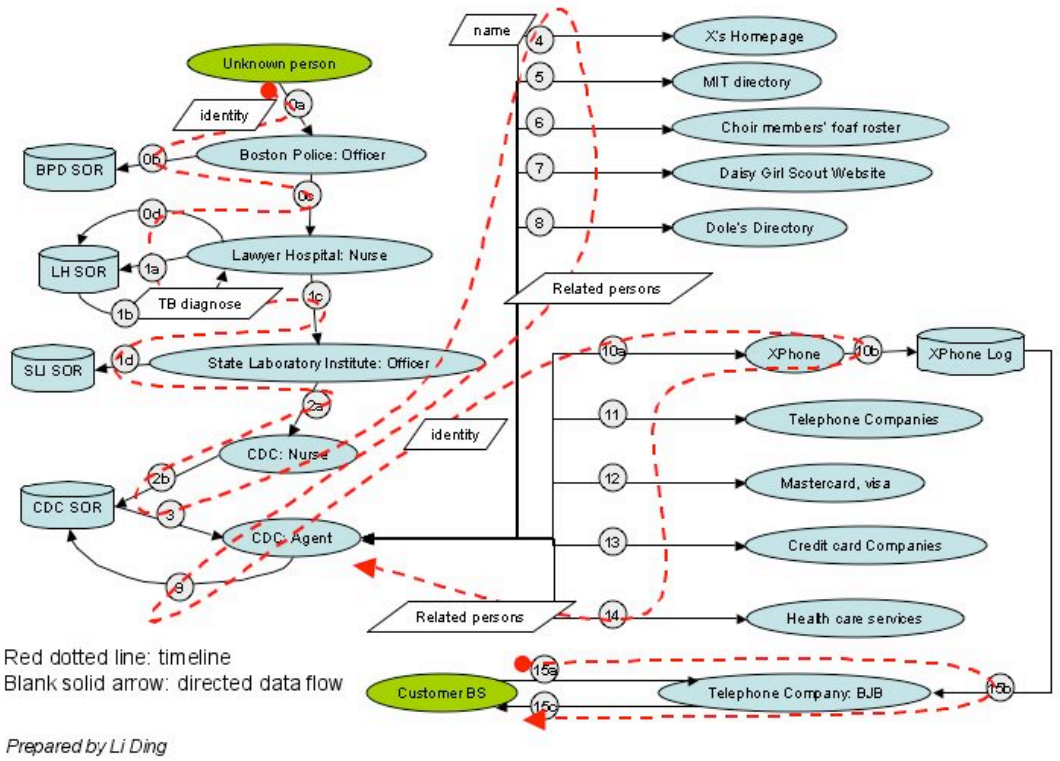
WSRI
web science research initiative

A Linked Data Social challenge

Public Health Investigation

Complex Scenario

Can we analyze data dependencies in order to track rule compliance and rule violation?



Security, Privacy and Trust

All economic, social and legal interactions are based on certain assumptions

Individuals can verify identities; can rely on the rules and institutions governing the interactions; and are assured that certain information will remain private.

On the Web: an environment where security, privacy and trust can be very difficult to monitor, verify and enforce. Will the Web grind to a halt as a result? Will ways be found to ensure that these basic features are present? Or will users of the Web find their own ways to cope with the absence of e.g., trust?

These questions call on a broad range of Web Science disciplines: to understand how individuals perceive trust and privacy when they use the Web; to see how concepts such as trust can be computationally represented; to develop the legal institutions needed to govern Web interactions. An understanding of the technology underlying security, the variables underlying trust and the extent of the privacy that Web users demand is clearly valuable in a number of industries.

Web Science

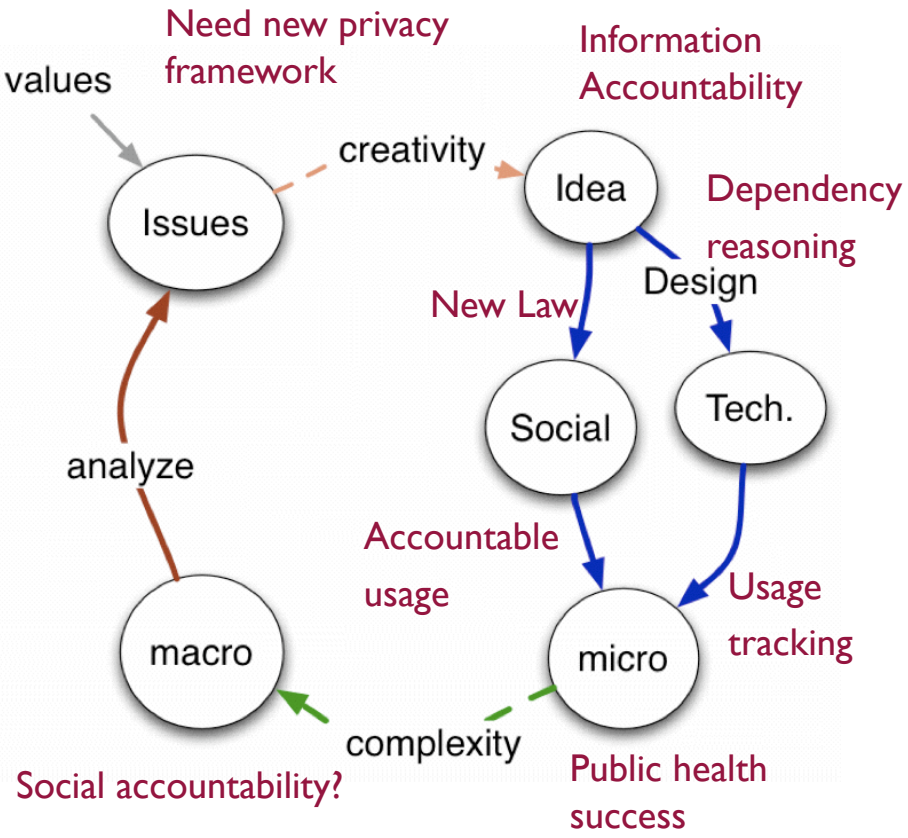
Information Accountability

New privacy framework

Technical architecture for usage tracking and policy explanation

Legal rules the concentrate on permissible/ impermissible use

Analysis of technical/social interplay



research / thought leadership / education

WSRI
web science research initiative

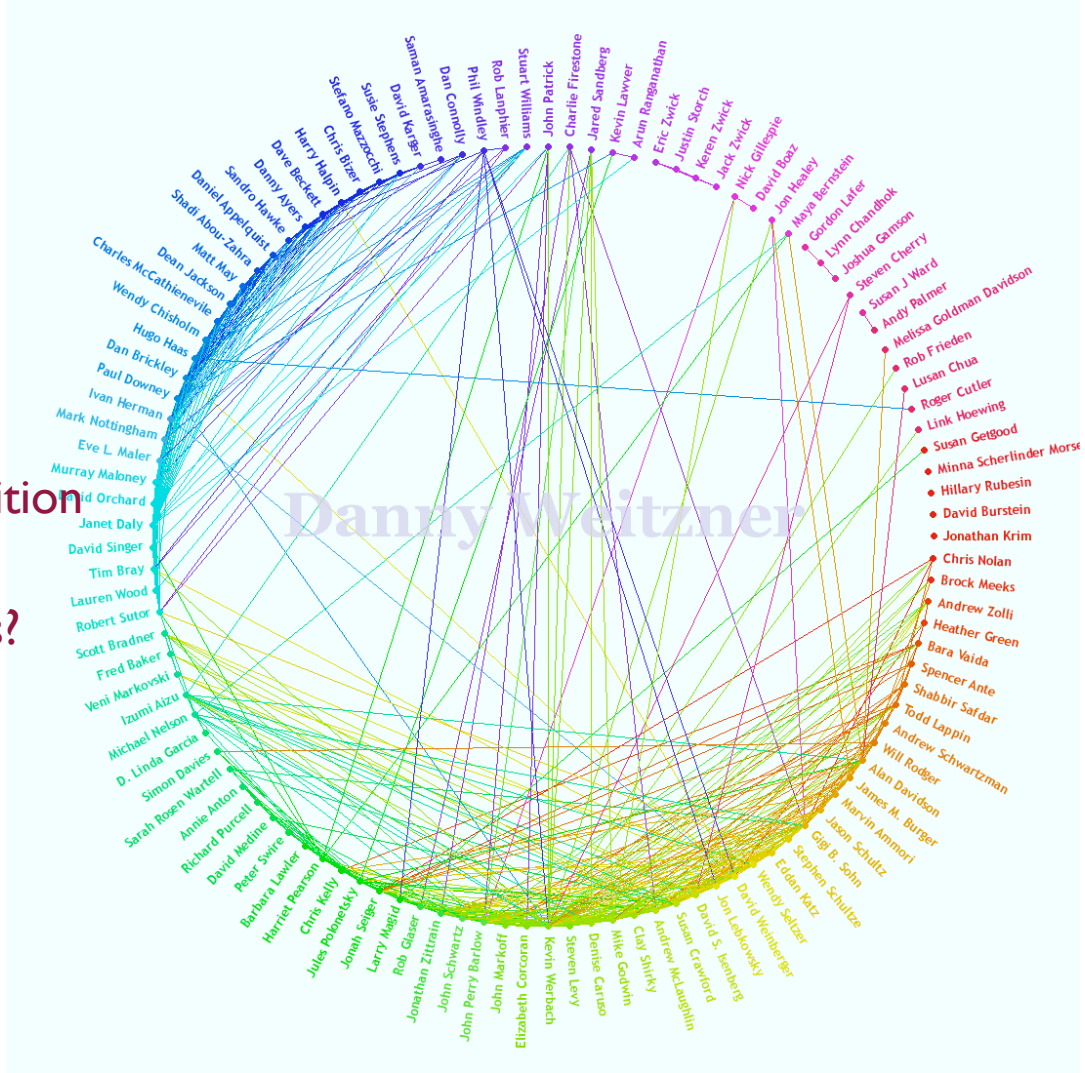
Web Science

Next social challenges…

What can we learn from composition of social networks?

What are the right privacy norms?

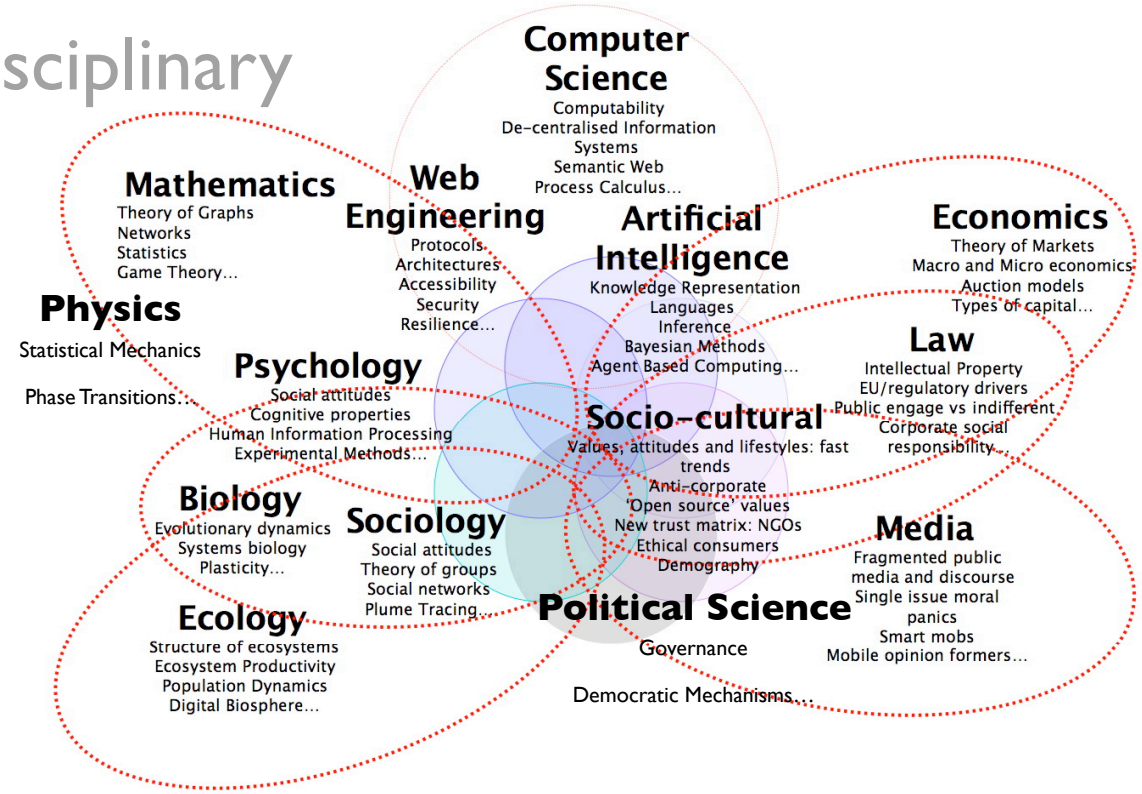
How will we arrive at social/legal consensus?



research / thought leadership / education

WSRI
web science research initiative

Web Science is
interdisciplinary



research / thought leadership / education

WSRI
web science research initiative

Computational Perspective

Web of linked data at a more fine-grained level - how we are to browse, explore and query such a Web at scale.

How do we support inference at a Web scale? What types of reasoning are possible? How is context represented and supported in Web inference?

How are concepts such as trust and provenance computationally represented, maintained and repaired on the Web?

As the Web has grown substantial amounts of it have become disconnected, atrophied or in others ways redundant. How are we to identify such necrotic and non-functional parts of the Web and what should be done about them?

Mathematical Perspective

How do we model the transient or ephemeral Web? How do we model this graph beneath the graph that is the Web?

What is the topological structure of the Web? Can connections always be established between its various parts, or do particular dynamic and time-dependent conditions create disconnected or sub- regions within it?

Given the huge numbers of searches performed simultaneously, the Web, at any given moment, will present a different structure to different users. It is a mathematical challenge to develop tools to describe this structure.

Social Sciences Perspective

How and why do people use newly emergent forms of the Web in the way that they do? What kinds of sociological and psychological concepts do we need to understand this? What implications does this have for our understanding of key sociological categories, e.g. kinship, gender, race, class and community, and vice versa? What implications does this have for our understanding of psychological constructs, e.g. personal and group identity, collaborative decision making, perception and attitudes.

How is the Web situated within networks of power and in relation to social inequalities? To what extent might the Web offer empowering political resources? How might the Web change further as new populations access it?

Economics Perspective

What are the economics of Web 2.0 (+)? What new economic issues are raised by the opportunities that Web 2.0 gives for users to generate content and share it in self-forming networks?

What are the economic forces that shape the formation of social networks on the Web? What is the relationship between the economic structure of the Web, its social and mathematical structure?

What are the commercial incentives created by the Web? What will be the industrial structure? Is the Web inherently prone to concentration, where a large part of the structure is owned and controlled by a small number of players? Or are there forces that will allow smaller scale operations to co-exist with large firms?

What are the economic arguments for and against open platforms in the Web? Should policy (economic and public) play any role in shaping or determining the openness of Web platforms?

Legal Perspective

Techniques for representing and reasoning over legal and social rules – what new tools need to be developed within legal theory to explore and understand the impact of law as a driver in shaping the Web development? Should law be a catalyst for change or merely reactive to it and how should it interact and respond to economic, social and technological influences?

Is the present intellectual property regulatory regime fit for purpose in the Web 2.0 (+) environment given that its legal principles were established in the offline world? What is content in the Semantic Web and what rights should attach to it particularly when much is likely to be "computer generated"?

Which technologies within the Web should the law ensure remain "open" rather than becoming the "property" of one or more commercial entities and what are the consequences of the choices available?

To what extent are the service providers going to become the legal gatekeepers for public authorities in terms of delivering their public policy objectives e.g. Web policing for what is judged to be "illegal and harmful content"?

What privacy issues arise in a Web environment of increasingly sophisticated information sharing?

UNIVERSITY OF Southampton

MIT

WSRI

web science research initiative

Activities

Outreach and Thought Leadership

- Refined Research Agenda with Sci Council (MIT Nov 2008)
- Influence on funding agencies – UK EPSRC (Digital Economy), European Commission, Singapore, China
- Workshops – WebEvolve2008 @WWW2008
- More workshops in the pipeline for 2009 – scholarship on the Web, accountability, critical infrastructure, WebEvolve2009?
- Web Science 2009, Athens, 18-20 March 2009

Education

- Doctoral Summer School, joint with the Oxford Internet Institute, July 2008
- Curriculum workshop, September 2008
- Curriculum wiki launched
- Another workshop or BoF in Athens @ WebSci'09
- Doctoral Training Centre Proposal

Fund-Raising

- "Below the line" research project funding at Southampton, MIT and RPI
- "Above the line" for WSRI's community building mission
  - NESTA (UK National Endowment for Science, Technology and the Arts)
  - Corporate Advisory Board
  - Donations and projects

WSRI Affiliation Activities

- WSRI Affiliated Labs (WAL's)
- Wider network of Web Science research groups
- Curriculum Development
- WSRI Ambassadors and Evangelists

WSRI Affiliated Web Science Labs

- Developing a network of Web Science Labs around the world
- Pursuing a coordinated programme of work
  - Research – annual meeting of research directors
  - Doctoral Summer Schools
  - Curriculum Development
  - Technology Transfer

research / thought leadership / education

WSRI
web science research initiative

Web Science

why this matters

- the Web matters
- an essential part of humanity
- understanding the Web is a major challenge as big as any other global cause

research / thought leadership / education

WSRI
web science research initiative