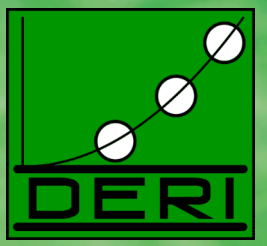


Advanced Studies in IT
CT433

Semantic Web - Information Integration with Semistructured and Web Data

Lecture 1.

© Copyright 2008 Digital Enterprise Research Institute. All rights reserved.

sfi science foundation ireland
fondúireacht eolaíochta éireann

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

High-level Introduction:

- **What is this lecture about?**
- Short History of the Web
- Beyond the current (static) Web
- Application areas
- What we will cover in the Lecture?

What is this lecture about?

- Information recovery, retrieval and integration from the Web.
- not only from the Web, but using Web technologies:
- at present: HTML, XML
- future: Semantic Web, Web services

High-level Introduction:

- What is this lecture about?
- **Short History of the Web**
- Beyond the current Web
- Application areas
- What we will cover in the Lecture?

The *World Wide Web* and its beginning…

- Hypertext
  - vision dates back to 1945: Vannavar Bush, *The Atlantic Monthly* called "As We May Think", the "Memex"
  - Ted Nelson, philosopher and IT pioneer, coined the term 'hypertext' in 1965, Xanadu project

    BTW: (WikiPedia) "Nelson hates the World Wide Web, the Internet, XML and all embedded markup, and regards Berners-Lee's work as a gross over-simplification of his own work. " ;-)
- Software
  - ENQUIRE (CERN, 1989)
  - Gopher
  - Mosaic (1993), Netscape, IE, etc.

Project ENQUIRE 1990 and WorldWideWeb

- Tim Berners-Lee 1990
- ENQUIRE used already terms like
  - "Universal Document Identifier"
  - Hypertext
- first browser and web server (WorldWideWeb, httpd)

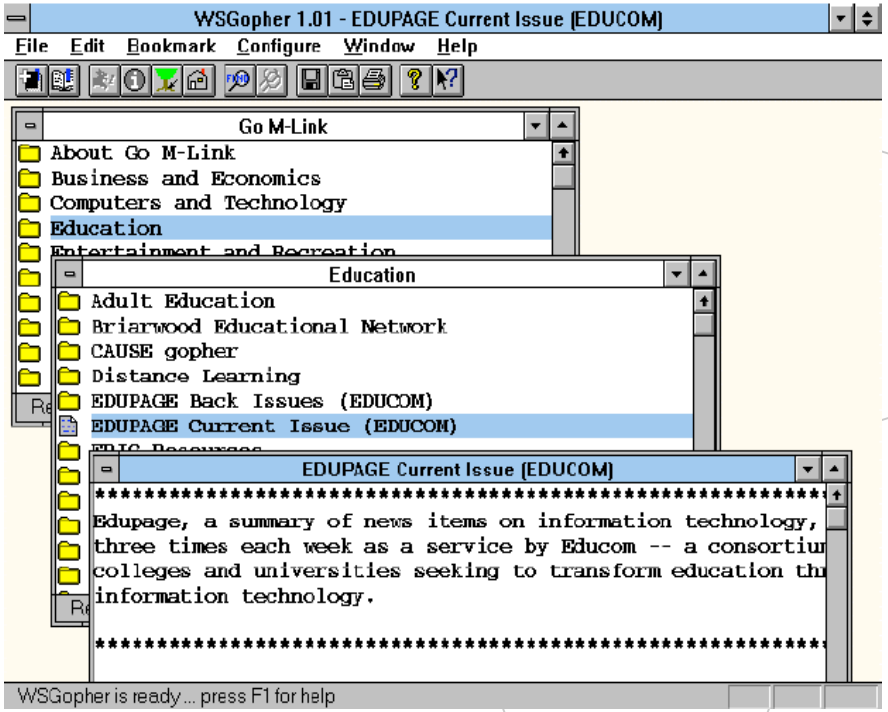
cf. http://www.w3.org/History.html

Gopher

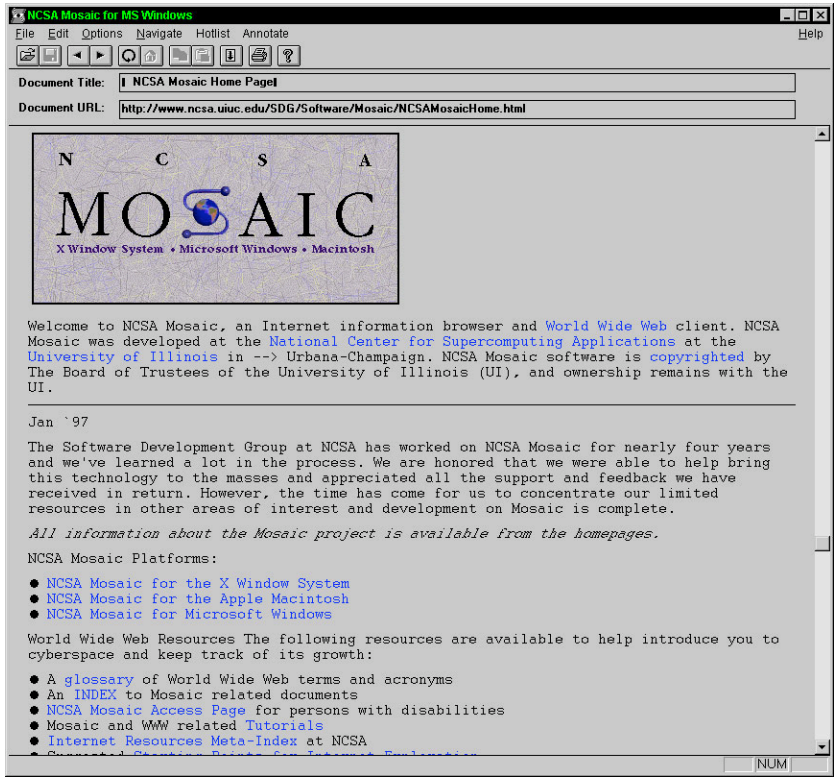
- First “Net Browser”, Univ. of Minnesota, 1991, no cryptic commands, menu-driven, network details completely hidden.
- Gopher Protocol:

```
gopher://
```

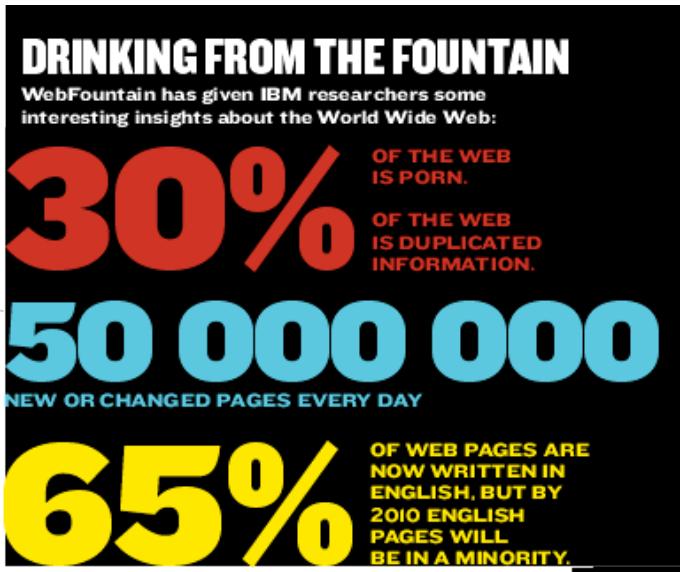
Hierarchies of application links, files, directories, phonebook server (X.500), grahics etc··· search indexing servers



Mosaic (1993), Netscape, IE, etc.



NCSA (National Center for Supercomputing Applications)
First Graphical browser…
V0.1 March 1993
Has lead to **500 million user**
more than 3 billion pages



First WWW Conference:

- **First International Conference on the World-Wide Web May 1994**
- **First W3 Consortium Meeting**: Dec 1994

Since then the W3C set up many important standard recommendations like XML (XML 1.0 Recommendation published on 10th February 1998), XML Schma, RDF, OWL, etc. .

http://www.w3.org

The current Web

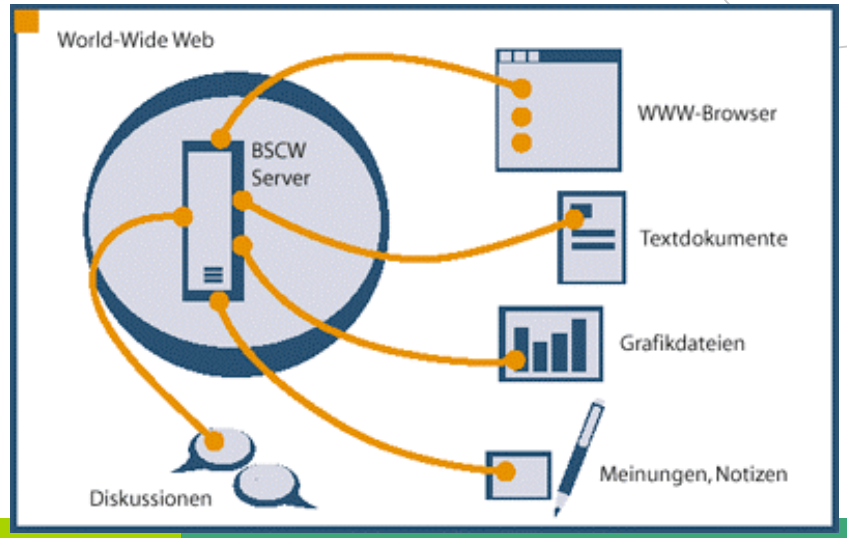
- Far from the pure Hypertext-Tool from the early days.
- The “biggest database” ever, but of the information is hidden in the “***deep web***” (dynamic data, behind forms, services, etc. approx. 500times bigger than the “surface web”!)
- Web-based applications heavily used in intranets as well, substituting classical applications. (platform-independence)
- eCommerce would not have become a reality at all without the web.
- Web opens new possibilities, but also overload of information

- BSCW (web-based Groupware)
- Wikis (community knowledge)
- Blogs (shared personal knowledge)

Example 1: BSCW (Basic Support for Collaborative Work)

- Example, for a web-based environment for collaborative work… http://bscw.gmd.de/
- a groupware system - suitable for small and medium enterprises as well as for world-wide operating companies
- shared workspaces
- No software installation, anywhere, anytime
- Combines Document Management, Calendaring, Discussion Groups
- Version control
- Self-organization of users

See: http://www.bscw.de/



Example2: Wikis

- New forms of KM not even existing before the Web.
- wiki “Simplest imaginable database”
- Empowered by Hypertextual linking
- Everybody can change everything
- Works anyway!
- Many different engines, nice Knowledge Management idea

For instance, see: http://moinmoin.wikiwikiweb.de/WikiSandBox
http://www.w3.org/2005/rules/wg/wiki/
http://www.wikipedia.org

Example 3: “Blogging, Weblogs”

- Web + Log = Blog
- Similar idea to Wikis, but more “sequential”
- Private or public logs which store information pieces in diary fashion.
- Could help in PIM (personal information management)! Blog what you do and find/link information
- People share/publish this information over the web.
- Others can publically check (but not change) this information.

some nice blogs:

http://dannyayers.com/
http://dig.csail.mit.edu/breadcrumbs/blog/2
http://inao.blogspot.com/
http://ivanherman.wordpress.com/tag/work-related/
http://danbri.org/words/

Examples of services over the Web:

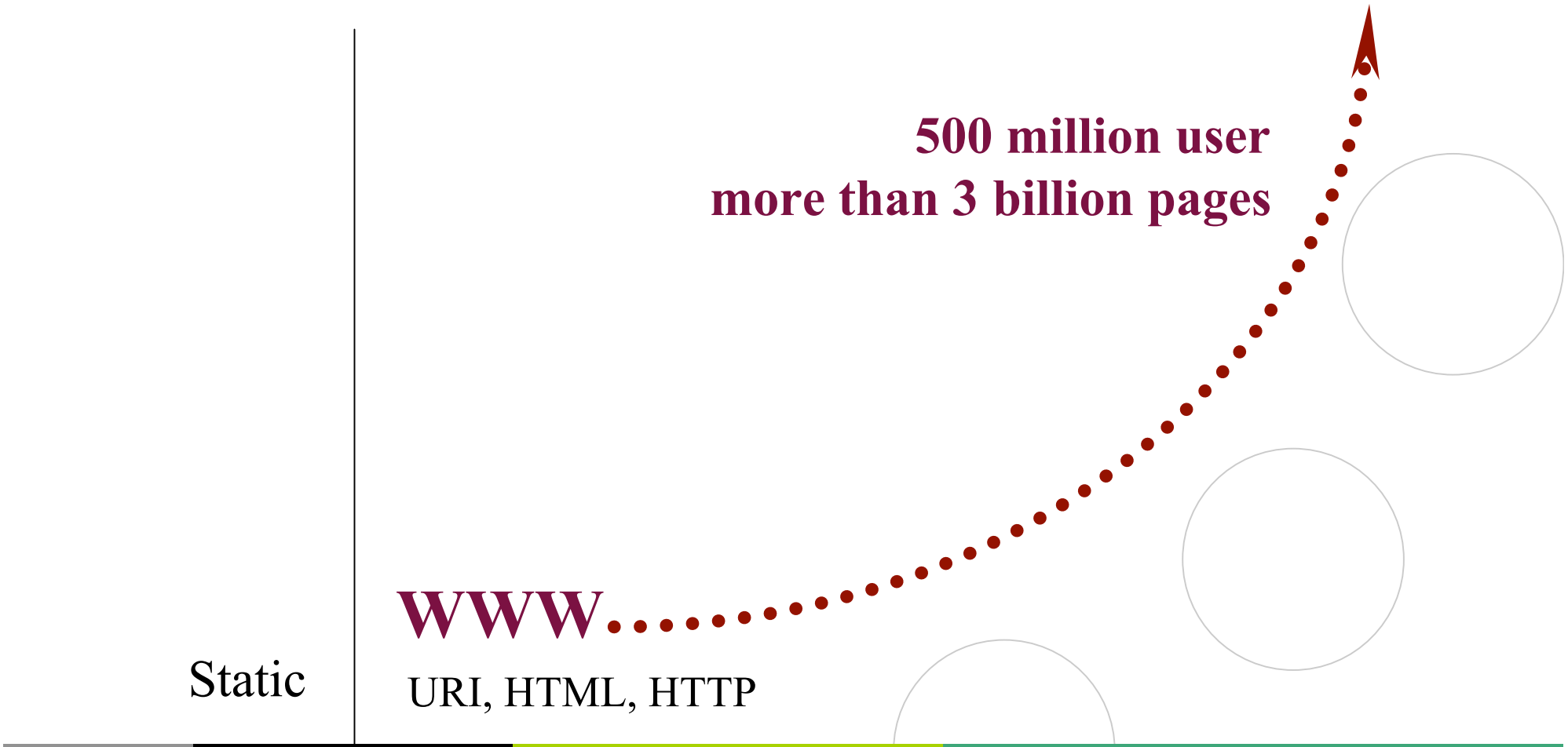
- ebay, forums, amazon, etc.etc.
- hotels, airlines, etc. offering their **service over the web**
- etc. …

  - http://www.dohop.com
  - http://www.amazon.com
  - http://www.expedia.com
  - …

High-level Introduction:

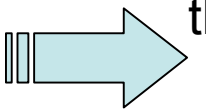
- What is this lecture about?
- Short History of the Web
- Beyond the current Web
- Application areas
- What we will cover in the Lecture?

Beyond the current Web: Next steps



The Vision

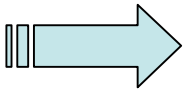
- The World Wide Web is a big and impressive success story, both in terms
  - of the amount of available information and
  - the growth rate of human users (50.000.000 changed or new pages/day)
- It starts to penetrate most areas of our daily life and business.
- This success is based on its **simplicity**



the restrictivenss of HTTP and HTML allowed software developers, information providers and users to make easy access of the new media helping it to reach a critical mass.

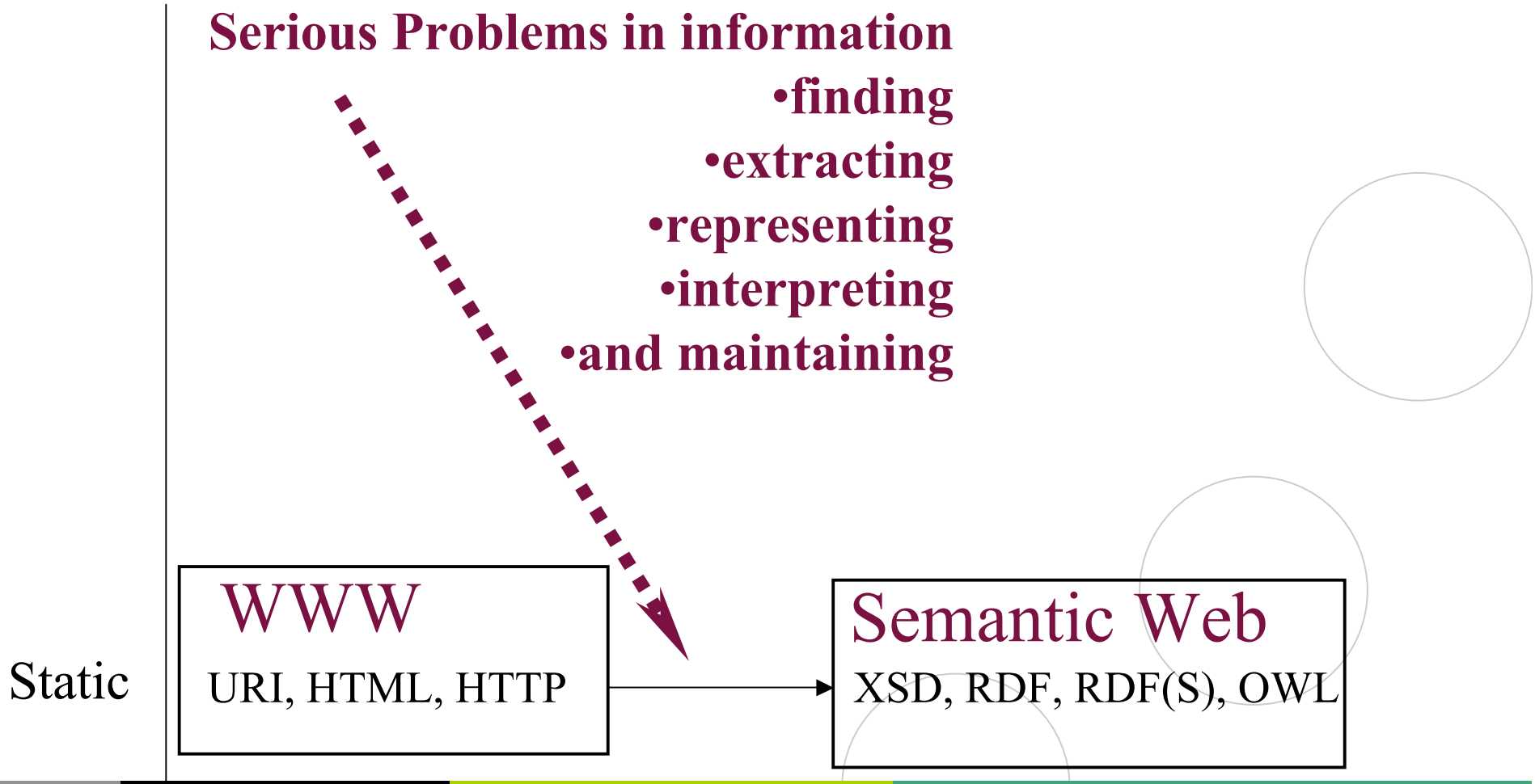
The Vision

- However this simplicity may hamper the further development of the Web



What we see currently is the very first version of the web and the next version will probably even bigger and much more powerful compared to what we have now.

Semantic Web



Semantic Web Technology

- Tim Berners-Lee has a vision of a Semantic Web which
  - has machine-understandable semantics of information, and
  - millions of small specialized reasoning services that provide support in automated task achievement based on the accessible information

The famous article…

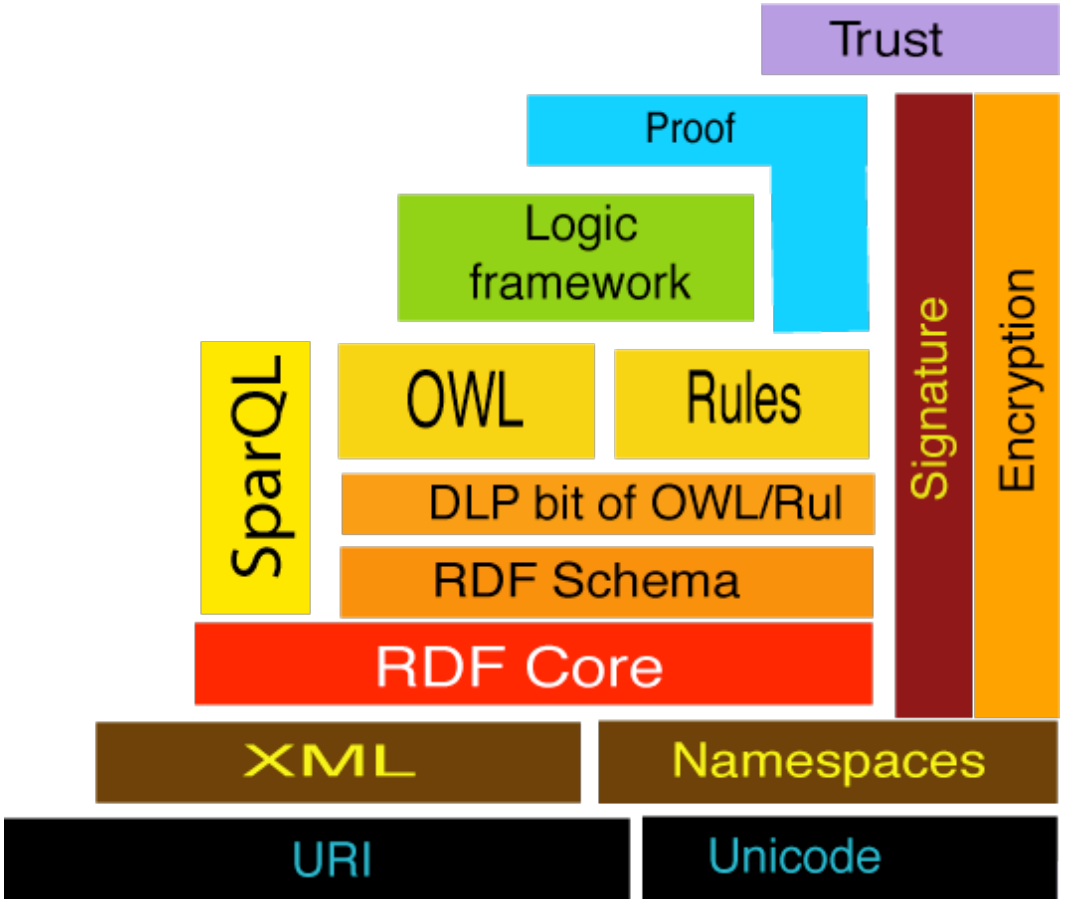
The Semantic Web (Scientific American Magazine - May, 2001)
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities By Tim Berners-Lee, James Hendler and Ora Lassila

The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other *local* devices that had a *volume control*. His sister, Lucy, was on the line from the doctor's office: "Mom needs to see a specialist and then has to have a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments." Pete immediately agreed to share the chauffeuring.

At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's *prescribed treatment* from the doctor's agent, looked up several lists of *providers*, and checked for the ones *in-plan* for Mom's insurance within a *20-mile radius* of her *home* and with a *rating of excellent* or *very good* on trusted rating services. It then began trying to find a match between available *appointment times* (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules. […]

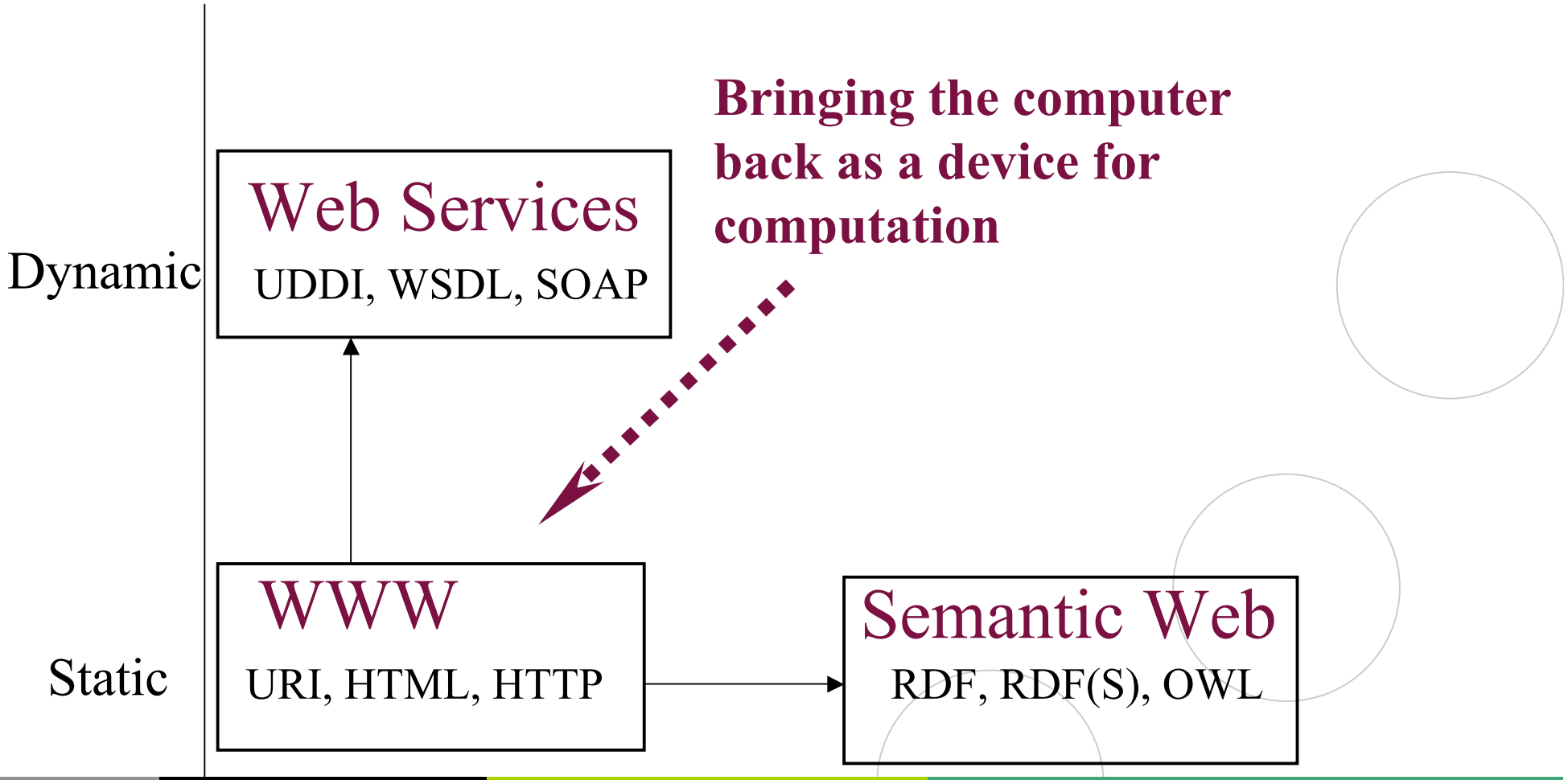
(The emphasized keywords indicate terms whose semantics, or meaning, were defined for the agent through the Semantic Web.) […]

It's a long way there …
Semantic Web - Architecture



http://www.w3.org/2005/Talks/0511-keynote-tbl/

Web Services



Web Services

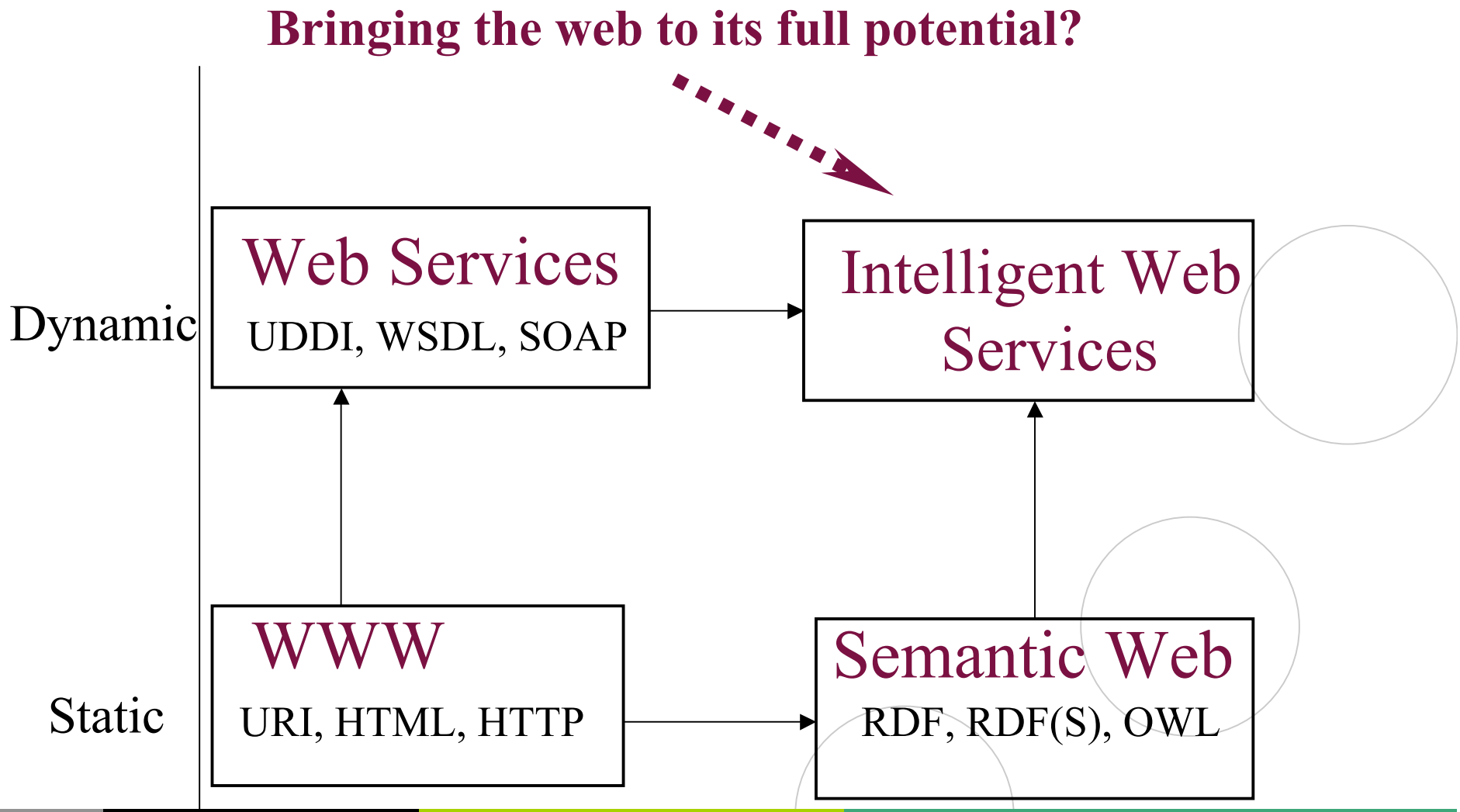
- Web Services connect computers and devices with each other using the Internet to exchange data and combine data in new ways.
- The key to Web Services is on-the-fly software creation through the use of loosely coupled, reusable software components.
- Software can be delivered and paid for as fluid streams of services as opposed to packaged products.

A (fictitous!)example: I regularly compare a list of online retailers and eBay automatically to determine whether I can offer cheaper than the cheapest eBay bid and automatically put an offer in case.
Make profit $$$!

Web Services

- "Web services" are an effort to build a distributed computing platform for the Web.
- Main tasks for making Web Services "machine processable":
  - Discovery: How can I find a service?
  - Composition: How can I combine services?
  - Execution: How can I execute services?
  - Monitoring: How can I monitor execution?

Semantic Web Service



The vision: Semantic Web Services

"Semantic differences, remain the primary roadblock to smooth application integration, one which Web Services alone won't overcome. Until someone finds a way for applications to understand each other, the effect of Web services technology will be fairly limited. When I pass customer data across [the Web] in a certain format using a Web Services interface, the receiving program has to know what that format is. You have to agree on what the business objects look like. And no one has come up with a feasible way to work that out yet **-- not Oracle, and not its competitors**..."

--- Oracle Chairman and CEO Larry Ellison

Semantic Web Services

- UDDI, WSDL, and SOAP are important steps into the direction of a web populated by services.
- However, they only address part of the overall stack that needs to be available in order to achieve the above vision eventually.
- There are many obstacles to achieve automatic web service discovery, selection, mediation and composition into complex services.

- The vision: combine Semantic Web technologies and Web Service techologies!

Short Summary

- The *Semantic web* is based on **machine-processable** semantics of data.
- It is based on new web languages such as XML, RDF, and OWL, and tools that make use of these languages.
- Web Services and Semantic Web Technologies shall help to achieve major improvements in core Data and service integration applications

Application Areas

- Knowledge Management
- Enterprise Application Integration
- eCommerce

Knowledge management

- What is KM
  - Knowledge is power!
  - To manage knowledge for the proper and efficient re-use

Why we need KM?

- Most of our work is information and knowledge based.
- Organizations compete on the basis of knowledge.
- Products and services are increasingly complex, endowing them with a significant information component.
- Reductions in staffing create a need to replace informal knowledge with formal methods.
- The amount of time available to experience and acquire knowledge has diminished.
- Early retirements and increasing mobility of the work force lead to loss of knowledge.
- The need for life-long learning is an inescapable reality.

Different views/disciplines

- Management:
  - learning organization
  - a cultural dimension of managing knowledge
- Economy:
  - competitive asset
- Artificial intelligence:
  - knowledge acquisition, knowledge engineering, knowledge-based system, computer understandable ontology
- Computer engineering:
  - knowledge management system, distributed, hypermedia tool

World Wide Web made KM the critical mass, also creates new challenges!

Two tracks in KM

- Management of Information
  - Knowledge = objects that can be identified and handled in information systems
- Management of people
  - KM are processes, a complex set of dynamic skills, know-how, etc, that is constantly changing.

Important: Knowledge needs structure! The Web has become the biggest source of Knowledge…

This is where (Semantic) Web Technologies come in!

Application Areas

- Knowledge Management
- Enterprise Application Integration
- eCommerce

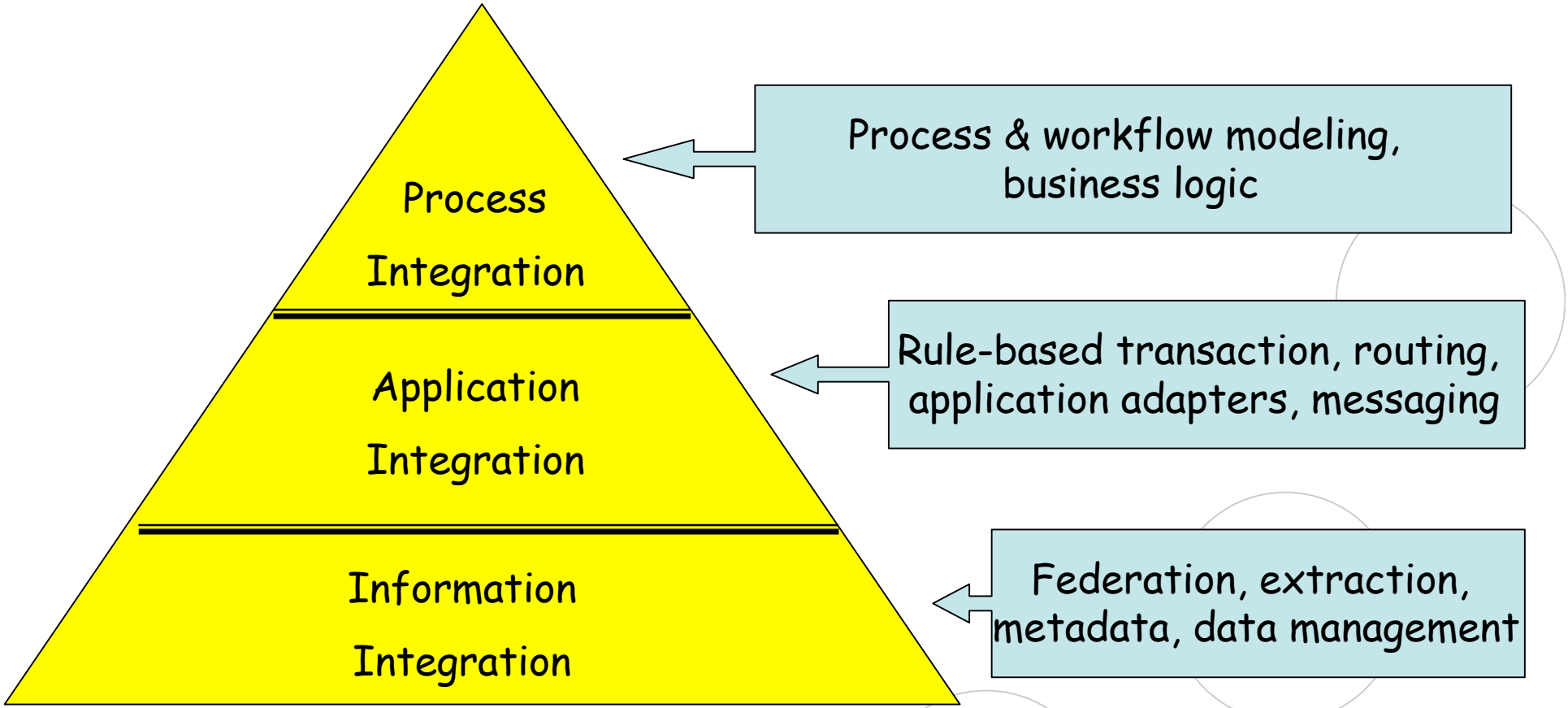
What is EAI

- The process of adapting a system to make applications work together when they would otherwise be incompatible.
- New applications in enterprises need to be integrated with legacy application because of previous investments
- Company mergers require integration of application from two different organizations

Why EAI

- Big market potential
  - integration middleware
  - application providers moving to integration infrastructures
  - Multiple, heterogeneous systems, no will possibilities to rebuild everything from scratch!
- Lack of integration of information costs:
  - Money
  - Reputation
  - Market share (through longer time-to-market)
  - Customers
- 30% of IT budget goes to integration

EAI Layers


Process
Integration
Application
Integration
Information
Integration
Process & workflow modeling, business logic
Rule-based transaction, routing, application adapters, messaging
Federation, extraction, metadata, data management

Integration - Example Semantic Differences:

Marketing

Person

P#	76798
Name	de Bruijn
FName	Jos
DName	Jos de Bruijn
BDate	1979-06-23
LSale	2001-04-07

?

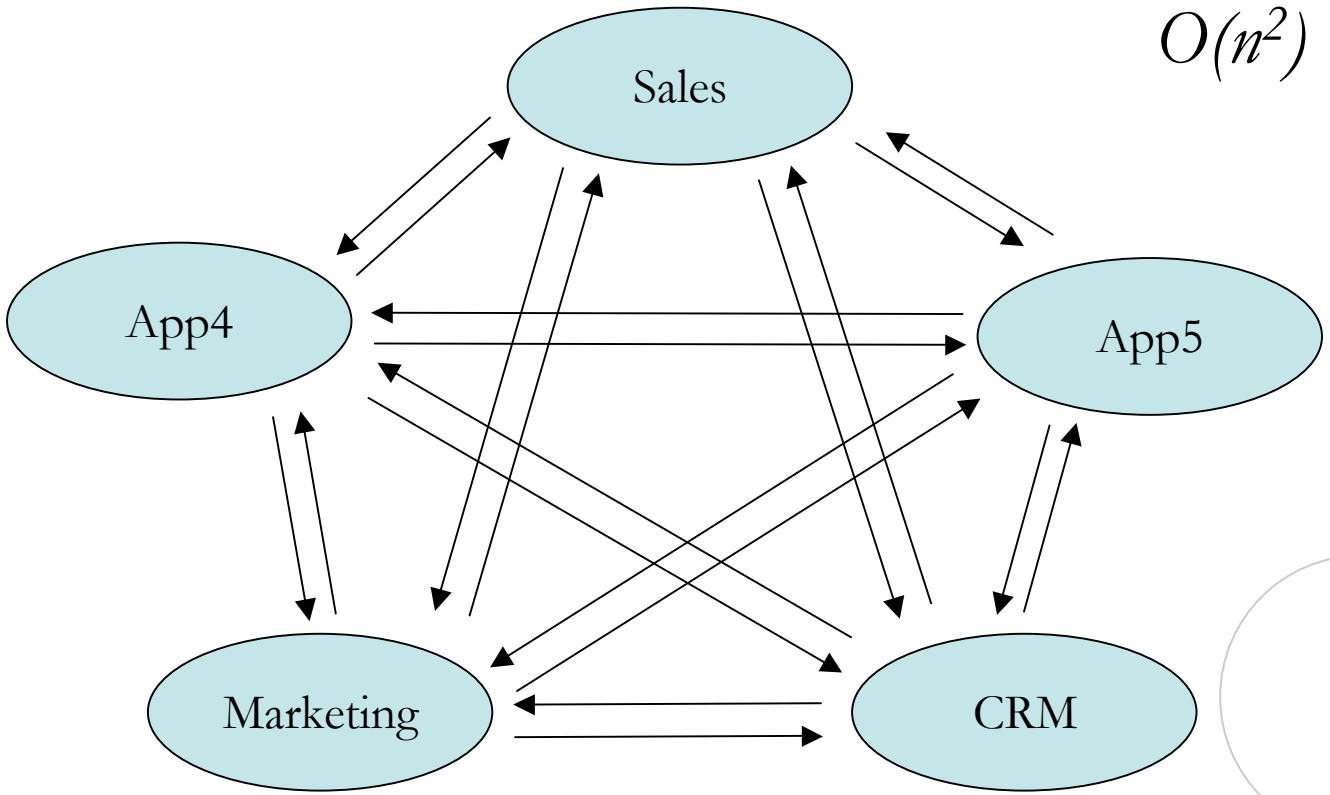
Sales

Customer

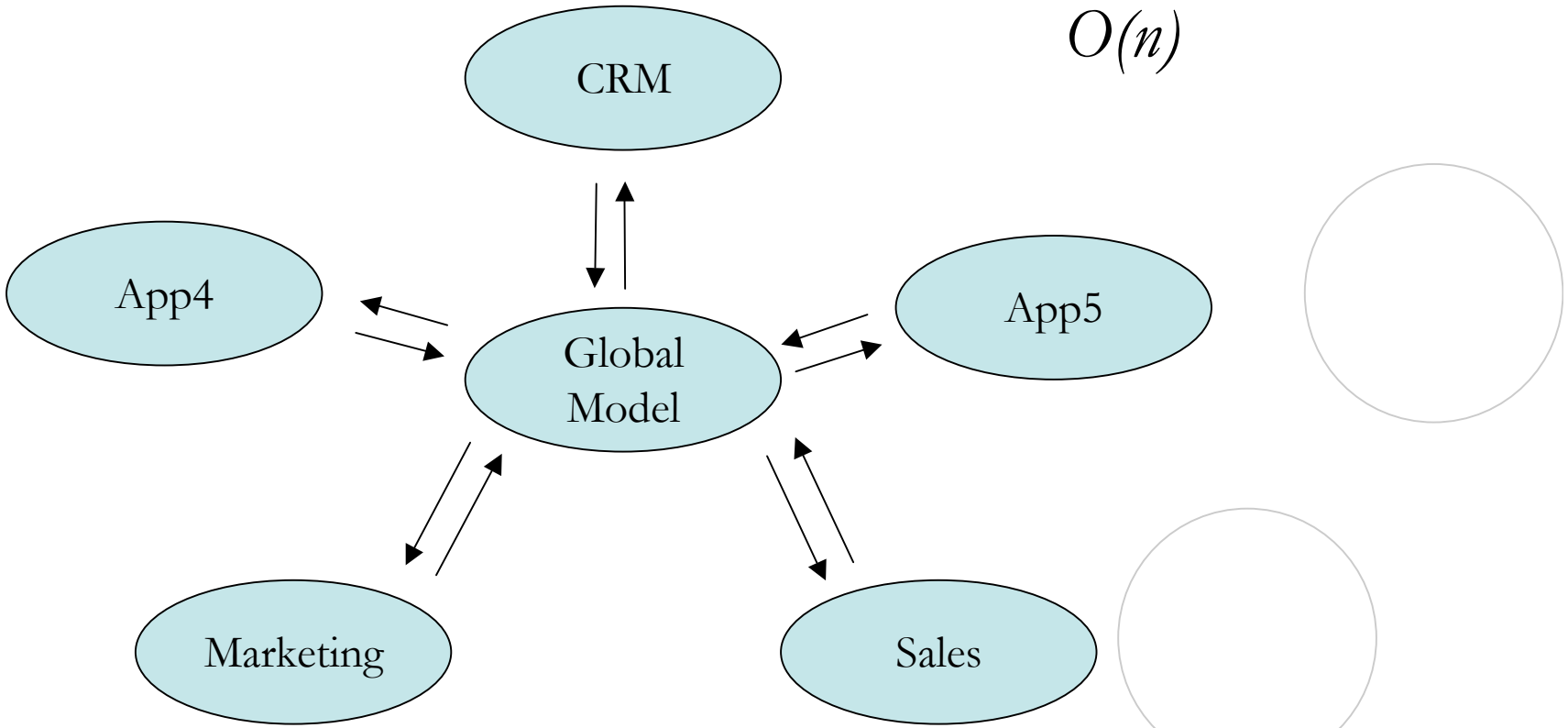
CustNr	43526
Name	Jos de Bruijn
Surname	de Bruijn
Initials	J
BDate	1979-06-23

	Syntax	Semantics
1.	distinct	equivalent
2.	equivalent	distinct
3.	equivalent	equivalent
4.	distinct	distinct

Information Integration Patterns (1):
Ad Hoc Integration: Mappings/Wrappers between all sources



Information Integration Patterns (2): *Global* Integration



Process & Application Integration

- Not only static data but dynamic applications/services need to be integrated!
- Web Services and emerging standards like WSDL, SOAP, UDDI offer means for integration, can help to solve the integration problem by providing common standards how applications can interact.
- Additionally, *Process integration* raises similar problems as the Information (Data) Integration Problem.

Summary: EAI needs Semantic Web Technologies and Web Services!

- On the syntax level: XML provides a common format.
- RDF and OWL (Ontologies) provide means to align semantic differences and define global schema information: Semantic Web technologies applied!
- Web Services are an emerging technology to make dynamic application integration happen!

Application Areas

- Knowledge Management
- Enterprise Application Integration
- eCommerce

eCommerce

- We expect that Enterprise Application Integration will be the major application of (Semantic) Web technology before it will take the next logical step:
- ➔ the integration of **several organizations**, i.e., eCommerce.

Content Management in E-Commerce

- WWW has magically changed our life
- E-commerce is one of the most important revolutions from the Web
  - B2C: 1% of the overall sales figures
    - small fraction but with the huge potential of user market
  - B2B: $600billion to $2.8 trillion

eCommerce

- eCommerce in business to business (B2B) is not a new phenomenon.
- However, the automatization of business transactions has not lived up to the expectations of the analysts propagandists.
- Establishing a eCommerce relationship requires a serious investment and it its limited to a predefined number of trading partners.

eCommerce

- Internet-based electronic commerce provides a much higher level of *openness, flexibility* and *dynamics* that could help to optimize business relationships.
- Anytime, anywhere, and anybody eCommerce provides completely new possibilities.

eCommerce

- However, enabling flexible and open eCommerce has to deal with serious problems.
- Heterogeneity in the *product*, *catalogue*, and *document* description standards of the trading partner.
- Effective and efficient management of different styles of description becomes a key obstacle for this approach.

Examples for existing WebServices and eCommerce

- amazon offers its services via SOAP, WSDL
- Xmethods www.xmethods.com lists publicly availabale web services, UDDI directory
- Google just turned off its web service interface ☹ Why?

First steps in directions of semantic search: www.froogle.com

Application Areas - Summary:

- What we have seen today:
  - The Web is a huge success story, enables new non-classical forms of telecooperation
  - The main problem is *integration*
    - KM, EAI, eCommerce

High-level Introduction:

- What is this lecture about?
- Short History of the Web
- Beyond the current Web
- Application areas
- What we will cover in the Lecture?

Main ingredients:

- XML
- Semantic Web Technologies (RDF, OWL, Rules, SPARQL
- Web Services

XML

- Tags define the semantics of the data

  <name>Axel Polleres</name>
- XML provides arbitrary trees (graphs) as data structures

  <person>

  <name>Axel Polleres</name>

  <phone>08154711</phone>

  </person>
- XML allows the definition of application-specific tags
- Provides a uniform, standard data-format for data exchange!

http://www.w3.org/XML/

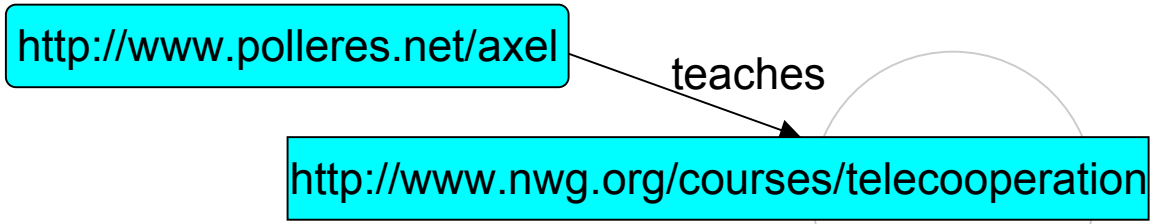
→ First Part of the Lecture

RDF

- XML provides semantic information as a by-product of defining the structure of the document
- That is, structure and semantics of documents are interwoven

- The Resource Description Framework (RDF) provides a means for adding semantics to a document without making any assumptions about the structure of the document and it provides pre-defined modeling primitives for expressing semantics of data.

```
<person>
<name>
Axel Polleres
</name>
<teaches>
<course><name>Telecooperation</name></course>
</teaches>
</person>
```



→ *Part II*

RDF Schema

RDFs provides a simple and basic modeling language for ontologies

- concepts
- properties
- is-a hierarchy and
- simple domain and range restrictions

be expressed in RDFs

Advanced ontology modeling need more, however, can be realized as a layer on top of RDFs

→ Part II

Ontology support - OWL

- Web Ontology Language - under development of W3C Web Ontology Working Group:
  - Offers:
    - Describing the structure of knowledge on the web
    - More accurate web searches
    - Intelligent agents
    - Reasoning (a bit of Logics!)

→ Part II

Web Services

- Yet another bunch of emerging standards ☺

  - **UDDI** provides a mechanism for clients to find web services. A UDDI registry is similar to a CORBA trader, or it can be thought of as a DNS service for business applications.
  - **WSDL** defines services as collections of network endpoints or *ports*. A port is defined by associating a network address with a binding; a collection of ports define a service.
  - **SOAP** is a message layout specification that defines a uniform way of passing XML-encoded data. In also defines a way to bind to HTTP as the underlying communication protocol. SOAP is basically a technology to allow for "RPC *over the web*".

→ Part III (if we get that far)

Recommended Reading:

The Semantic Web (Scientific American Magazine - May, 2001)
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities
By Tim Berners-Lee, James Hendler and Ora Lassila
http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

The Semantic Web In Action (Scientific American Magazine - December, 2007)
Corporate applications are well under way, and consumer uses are emerging
By Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann and Susie Stephens
http://www.sciam.com/article.cfm?id=the-semantic-web-in-action

Thanks!