From Web to Semantic Web
From Internet to Web 2.0

December 1969

June 1970

March 1972

July 1977
First Generation: The Internet
Computer Centered Processing

- How did the user get the information?

1. open terminal
2. connect to remote computer
3. retrieve file system data from remote computer
4. download file from remote to local computer
5. read file on local computer

Problem:
- Information access requires expert knowledge
- Information access is expensive...
- Information retrieval is very expensive...

Hasso Plattner Institute
The World Wide Web was born at the European Nuclear Research Center, CERN in 1990...
Second Generation: The Web
Document Centered Processing

- How did the user get the information?

Advantages:
- No expert knowledge required.
- Simple information access
- Information retrieval via search engines
Vannevar Bush proposed the first Hypertext-System „MEMEX“ in 1945.

Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (LIFE 19(11), p. 123).
From Web 1.0 to Web 2.0
Web Content and Applications are Changing

Information Consumption

Information Production

Interactive Participation

1968: IRC
1978: BBS
1979: Usenet

- Social Bookmarking
- Podcasting
- Social Networking
- RDF/RSS
- Blogs
- Blogger
- Movable Type
- Napster
- Gnutella
- BitTorrent
- Wikis
- Wikipedia
- Wildnews
- ICQ
- AIM
- Jabber
- Forums
- Yahoo! Groups

Lotus Notes
IMDb moves to WWW
Yahoo!
eBay
GeoCities
Slashdot
OhmyNews
Drupal CMS
Napster
Google News
Sidekick
Skype
Gnu Media
Flickr

Semantic Web Technologies, Dr. Harald Sack, Hasso Plattner Institute, University of Potsdam
• 웹 2.0의 개념 (~ ~ 2004)

1. 웹은 플랫폼이다.
   The Web As Platform
2. 집단지성을 활용한다.
   Harnessing Collective Intelligence
3. 데이터가 차별화의 열쇠다.
   Data is the Next Intel Inside
4. 소프트웨어 배포 주기란 없다.
   End of the Software Release Cycle
5. 가볍고 단순하게 프로그래밍한다.
   Lightweight Programming Models
6. 소프트웨어는 PC에 익매이지 않는다.
   Software Above the Level of a Single Device
7. 사용자들에게 더 많은 편리함을 제공한다.
   Rich User Experiences
Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform."

-- Tim O'Reilly, 2003
flickr.com: 사용자가 사진을 올리고, 올린 사진에 대해 태깅을 달고 다른 사용자는 댓글을 달아 검색을 용이하게 하기 위한 서비스이다.
del.icio.us: 온라인 상의 북마크 사이트이다. 사용자는 온라인 상에서 북마크를 할 수 있으며 또한 다른 사용자들은 그 사이트가 얼마나 많이 저장 되었는지, 어떤 사용자들이 저장을 했는지 등에 대한 정보를 얻을 수 있다.
Wikipedia: 온라인 백과 사전으로 누구나 글을 쓰거나 수정/보완 할 수 있음. 다른 사람의 위키에 첨가, 삭제, 변형이 가능. 하루 평균 870만회 조회수. 전 세계 네티즌에 의해 200여개 언어로 작성
Newmap: 뉴스를 읽기 위해 회원 가입이나 광고를 볼 필요가 없을 뿐 아니라, 뉴스를 선택하여 보기 위해 사이트를 이동할 필요도 없는, 즉 모든 뉴스를 모아놓는 서비스이다.
Google Adsense: 자신의 웹 페이지나 Blog에 광고를 올릴 수 있고 클릭 횟수에 따른 수익금을 받을 수 있다. 웹 컨텐츠와 관련된 광고가 동적으로 디스플레이되며, 광고 장소 제공자는 클릭 횟수에 따른 수익금을 얻는다.

레알

월드컵 끝난 마당에 레알(Real)을 보게 된 이유는 Firespotting의 추천 때문이었다.

5년간의 재작년, 5개국 로케이션, 5가지 롤러바스 이야기가 멋있게 세계 최고의 축구팀 레알 마드리드(Real Madrid)의 드라마와 함께 달겨졌다. 롤러바스 드라마의 대표적인 리브 맥후如果玩家(Love Actually)를 생각하면 들리지 않다. 단순히 레알 마드리드의 축구 드라마를 통해 부가 기성 6가지 에피소드가 차지하는 비중 또한 결코 적지 않다.

하지만 이 영화의 주인공은 레알 마드리드이고 "레알 마드리드라는 축구팀이 이렇게 많은 팬들의 삶과 그들의 영광에 영향을 미쳤다는 Firespotting"에 대한 이야기다. 특히 후반부 바르셀로나(FC Barcelona)와 경기장면은 축구가 얼마나 맞잡은 스포츠인지 느끼기에 충분하다.

또다음을 꺾고 축구영화로 Seo님의 추천한 "앞드리 되어야 할 축구영화"도 배경을 수 있다.

어쨌든 공이 이 포스팅을 남긴 이유는 다음과 있다. 바로 영화중간, 축구에 전혀 관심없던 역사선생님이 레알 마드리드에 관심을 갖기 시작하는 모습에 휩쓸리는 한편 때로.
creativecommons.org: License 계약을 체결하기 위한 일종의 표준 약관과 같다. 저작재산권자는 저작물의 이용을 허락할 수 있고, 허락을 받은 자는 "허락 받은 이용방법 및 조건의 범위 안에서" 저작물을 이용할 수 있다.
Web2.0 특징: 참여를 통한 컨텐츠 생산

UCC(User-Created Contents)
Web 2.0: 컨텐츠의 생산과 유통
Web2.0 vs Semantic Web [1/2]

- **Semantic Web**
  - 기존의 Web Contents에 의미를 저장 & 제공하여 사람과 컴퓨터가 좀더 협력을 잘 할 수 있는 웹 (by Tim Burners Lee)
  
  - 컴퓨터끼리 정보해석력을 높여서 자동화 처리가 향상되는 웹

- **Web 2.0**
  - Web 2.0은 Semantic Web을 경제적 관점에서 보는 표현
  
  - Semantic Web에 의해 개발된 기술을 이용, 좀 더 효율적이고 지능적인 웹생활의 구현이 Web 2.0의 목표

- Business Sector ➔ Web 2.0
- Academic Sector ➔ Semantic Web

- Web 2.0 ~~ Semantic Web ~~ 차세대 웹 ~~ 미래의 웹
Web2.0 vs Semantic Web [1/2]

- **Web2.0**: 풍부한 Web-contents를 위해서 사용자의 참여를 적극적으로 유도하는 개방적 체계 (User Created Contents)
  - 오래된 기술: Blog, P2P
  - 최근의 기술: RSS, Tagging

- **Semantic Web**: 기존의 Web에서 표현되는 정보의 의미 (semantics)를 새로운 기법으로 강력하게 표현하고 저장 (Web Intelligence)
  - Ontology 표현기술: RDF, OWL
  - 지능형 검색과 지능형 웹서비스를 가능
  - 컴퓨터끼리 대화하고 자동화된 지능형 웹을 추구
Web Intelligence의 예 (1/2)

- 기존의 web:
  - 1) 환자가 검색 엔진에서 치과를 검색
  - 2) 자신이 거주하고 가까운 치과의 홈페이지를 찾음
  - 3) 치과의 진료 스케줄을 확인하고 자신과 시간이 맞을 경우 예약
  - 예약을 하기까지 다수의 반복 작업 필요
Web Intelligence의 예 (2/2)

- **Semantic web**으로 다음의 정보가 구축된 상태라면
  - 환자의 개인 스케줄, 각 치과의 위치, 진료 과목, 진료

(1) 환자는 **software agent**에게 예약 요청

(2) 각 병원의 홈페이지의 내용이나 구조가 다르더라도 **software agent**가 환자와 치과의 시멘틱웹 데이터를 분석, 환자의 시간과 위치에서 진료 가능한 치과를 예약해 줌
Super Challenge!

Intelligent Information Retrieval
(지능형 정보 검색)

How can we find something in the Web?

current solution:

Google
The Web is really big...

- ca. $25 \times 10^9$ indexed documents in search engines
  (TNL Blog: Google has 24 billion items index, considers MSN search nearest competitor, September 2005)

- Web Crawler: $> 10^{12}$ (1 trillion) documents
  (The Official Google Blog: We knew the Web was Big....., Juli 25, 2008)

- Google Search Index Caffeine comprises
  ca. 100 Million Gigabytes i.e. $10^{17}$ Byte

- DeepWeb (Darkweb) estimated to be about 550 times bigger than Surface Web
  (Bergman, 2001)
WITH NO SIGNS OF SLOWING, THE DATA Keeps GROWING

There are just some of the more common ways that internet users add to the big data pool. In truth, depending on the niche of business you’re in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

- The global internet population grew 6.59 percent from 2010 to 2011 and now represents 2.1 billion people.
- Flickr users add 3,125 new photos every minute.
- Instagram users share 3,600 new photos.
- Tumblr blog owners publish 27,778 new posts.
- Twitter users send over 100,000 tweets.
- Apple receives about 47,000 app downloads.
- 571 new websites are created every minute.
During 2008, the number of things connected to the Internet exceeded the number of people on earth.

By 2020 there will be 50 billion.
Information in the WWW

- what is *important* and how do you know?
- what is information, what is *advertisement*?
- what does the information *mean*?
- how *credible/trustworthy* is the information?
- what belongs *together*?
- what is *redundant*?

- Humans have *contextual knowledge, world knowledge* and *experience* to solve the problem
- The Web is supposed to be used by humans
- The Web is based on the **markup language HTML**

- HTML describes
  - how information is **presented**
  - how information is **linked**
  - but **not** what the information means

WHAT IS THE MEANING?
Semantics

Semantics (Greek σεμαντικός = pertains to the character; the study of meaning) is part of the linguistics focussed on

- Sense and
- Meaning

of language or symbols of language. It is the study of interpretation of signs or symbols as used by agents or communities within particular circumstances and contexts.

Semantics asks, how sense and meaning of complex concepts can be derived from simple concepts based on the rules of syntax.

The semantics of a message depends on its context and pragmatics.
Syntax (Greek σύνταξις = Arrangement, Ordering) as in grammatics denotes the study of the principles and processes by which sentences are constructed in particular languages.

- In **formal Languages**, syntax is just a set of rules, by which well-formed expressions can be created from a fundamental set of symbols (alphabet).
- In **computer science**, syntax defines the normative structure of data.
Context (lat. contextus = interweaved) denotes the surrounding of a symbol (concept) in an expression resp. its relationship with surrounding expressions (concepts) and further related elements.

Contexts denotes all elements of any sort of communication that define the interpretation of the communicated content, as e.g.,

- **general contexts:**
  - place, time, interrelation of action in a message
- **personal or social contexts:**
  - relation between sender and receiver of a message

Pragmatics (greek. πράγμα = action) reflects the intention by which the language is used to communicate a message.

In linguistics pragmatics denotes the study of applying language in different situations. It also denotes the intended purpose of the speaker. Pragmatics studies the ways in which context contributes to meaning.
Experience considers all information that you have learned and put in context with the world you are living in.

For successful communication,
- information has to be correctly transmitted (Syntax)
- the meaning (Semantics) of the transmitted information must be interpreted correctly (= understanding)
- understanding depends on
  - the context of both sender and receiver and
  - the pragmatics of the sender
- the context of sender and receiver depends on
  - the experience (knowledge of the world) of both sender and receiver
Meaning

Concept

Symbol

Object

sender

receiver

symbolizes

refers to

stands for

"Jaguar"

Pragmatics
How do we get the meaning in the Web...?

Problem 1: Information Retrieval

- traditional keyword-based search leads to many not relevant results
  - different meanings
  - polysemy
  - different contexts
Problem 1: Information Retrieval

- traditional keyword-based search does not find all results
- synonyms and metaphors
- missing context definition
Information Extraction

- can only be solved 'correctly' by a human agent
- heterogeneous distribution and order of information
- Software agent does not have sufficient
  - knowledge of contexts
  - world knowledge and
  - experience
- to solve the problem

Problem 2: Information Extraction
*implicit knowledge*, i.e. information does not have specified explicitly, but must be derived via logical deductions from available information.

**Problem 2: Information Extraction**
Problem 3: Maintenance

- the more complex and voluminous a website, the more complicated is the maintenance of the only weakly structured data.

- Problems:
  - syntactic and semantic (link) consistency
  - correctness
  - timeliness
Adaption of the presented information content to personal requirements

Problems:
- from where do we get the required (personal) information?
- personalization vs. data security

Problem 4: Personalization
The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help... “

Tim Berners-Lee, Semantic Web Roadmap, Sept 1998
From the World Wide Web to the Web of Data

Precondition:
- Content can be **read and interpreted correctly** (≠ „understood“) by machines

Natural Language Processing
- Technologies of **traditional Information Retrieval** (Search Engines)

Semantic Web
- Natural language web content will be **explicitly annotated with semantic metadata**
- Semantic metadata encode the **Meaning (Semantics)** of the content and can be read and interpreted correctly by machines
"Understanding" Content on the Web

Text: "Why snub me, Fab?"

Entity Mapping Disambiguation

- Fab ?
- Fab ?
- FAB ?
- Fab Melo ?
- Fabrizio Moretti ?
- Fabio Capello ?
- a detergent
- a Canadian gay magazine
- the Film Advisory Board
- Federación Atlética de Bolivia
- Brazilian Basketball Player
- Drummer of The Strokes
- Manager of the UK national Football team

Disambiguation
solution of linguistic ambiguities
Understanding Content on the Web

Text: "Why snub me, Fab?"

Entity Mapping

- Fabio Capello
- Soccer Manager

The Meaning (Semantics) of entities and classes must be defined explicitly.
The Meaning (Semantics) is expressed with the help of appropriate knowledge representations (Ontologies).
The Meaning (Semantics) is expressed with the help of well suited knowledge representations (Ontologies).
"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"

The Semantic Web - a Web of Data

- The Meaning of Information (Semantics) is made explicit by formal (structured) and standardized knowledge representations (Ontologies).

- Thus it will be possible,
  - to process the meaning of information automatically
  - to relate and integrate heterogeneous data
  - to deduce implicit (not evident) information from existing (evident) information in an automated way

The Semantic Web is kind of a global database that contains a universal network of semantic propositions.
Semantic Web
...towards an „intelligent“ Web

The Semantic Web Technology Stack
(not a piece of cake...)

Most apps use only a subset of the stack
Querying allows fine-grained data access
Standardized information exchange is key
Formats are necessary, but not too important
The Semantic Web is based on the Web

Linked Data uses a small selection of technologies
Semantic Web
...towards an „intelligent“ Web

URI - Uniform Resource Identifier

Fabio Capello → http://dbpedia.org/resource/Fabio_Capello
Semantic Web
...towards an „intelligent“ Web

RDF Version of Wikipedia
http://de.wikipedia.org/wiki/Fabio_Capello

http://dbpedia.org/resource/Fabio_Capello
Semantic Web
...towards an "intelligent" Web

http://dbpedia.org/resource/Fabio_Capello

About: Fabio Capello
An entity of Type: agent, from Named Graph: http://dbpedia.org

Fabio Capello es un entrenador que ha dirigido equipos de fútbol en varios países y clubs diferentes. Nació en 1943 en Italia y comenzó su carrera como entrenador en la Escuela de Fútbol de Roma. Ha dirigido diferentes equipos de fútbol en diferentes países, incluyendo el Real Madrid, el AC Milan y el Juventus. Ha ganado numerosos títulos en sus equipos, incluyendo la Copa Italia, la Liga de Campeones de la UEFA y la Supercopa de Europa.

Valores
- Cumplimiento: Fabio Capello es conocido por su ética y disciplina. Es un entrenador que siempre busca el mejor rendimiento de sus jugadores.
- Respeto: Fabio Capello tiene una gran reverencia por el fútbol y su trabajo en el deporte, y siempre busca el mayor respeto posible.

Fabio Capello ha dirigido equipos internacionales en diferentes países, incluyendo Italia, Inglaterra y Francia. Ha sido un entrenador muy respetado y admirado por su estilo de juego y su capacidad para dirigir equipos exitosos en diferentes países.

DBpedia
Semantic Web
...towards an „intelligent“ Web

http://dbpedia.org/resource/Fabio_Capello

:Fabio_Capello dbpp:birthPlace :San_Canzian .
:Fabio_Capello rdfs:type dbpo:SoccerManager .
:Fabio_Capello rdfs:type dbpo:Person .

RDF Resource Description Framework

RDF Tripel

:Fabio_Capello ↓ RDF Subject
 rdf:type ↓ RDF Property
dbpo:SoccerManager ↓ RDF Object
Semantic Web
...towards an „intelligent“ Web

http://dbpedia.org/ontology/SoccerManager
dbpo:SoccerManagerrdf:typeowl:clas

dbpo:SoccerManager rdfs:subClassOf dbpo:Person

dbpo:SoccerManager rdfs:label "SoccerManager"

dbpo:birthPlace rdf:type rdf:Property

dbpo:birthPlace rdfs:domain dbpo:Person

dbpo:birthPlace rdfs:range dbpo:Place

dbpo:birthDate rdf:type rdf:Property

dbpo:birthDate rdfs:domain :Person

dbpo:birthDate rdfs:range xsd:date

RDF Schema

SoccerManager

↓ is subclass of

Date ← birthDate

Person ← birthPlace → Place
Semantic Web
...towards an „intelligent“ Web

1946-06-18

entities
birthDate

classes

is a

LivingPeople ∩ DeadPeople = ∅

logical constraint

∀x. ∃y. hasDeathDate(x, y) ∧ Person(x) ∧ Date(y) → DeadPeople(x)

(W3C, OWL)

(rules)

+ Rules
Look for all Soccer Players that have scored more than 10 goals as member of a national team.

```sparql
PREFIX dbpo: <http://dbpedia.org/ontology/>
PREFIX dbpp: <http://dbpedia.org/property/>

SELECT DISTINCT ?I ?I2 ?g
FROM <http://dbpedia.org>
WHERE {
  ?s rdf:type dbpo:SoccerPlayer .
  ?s dbpp:nationalteam ?o .
  ?s rdfs:label ?I
    FILTER langMatches( lang(?I), "EN" ) .
  ?s dbpp:nationalgoals ?g
    FILTER(?g>10).
  ?nat rdfs:label ?I2
    FILTER langMatches( lang(?I2), "EN" ) .
}
ORDER BY DESC(?g)
```
Look for all Soccer Players that have scored more than 10 goals as member of a national team.

Problem: Data Quality....
Semantic Web Technologies
Applications on the Web of Data

Linked Data

- Linked Open Data (LOD) denote publicly available (RDF) Data in the Web, identified via URI and accessible via HTTP. Linked data link to other data via URI.

The 'Web of Data'
- >31 billion facts
- >500 million links
  (Oct. 2011)
What is so special about the BBC Music Website?

- Information is dynamically aggregated from external, publicly available data (Wikipedia, MusicBrainz,...)
- No Screen Scraping
- No specialized API
- Data available also Linked Open Data
- Data access via simple HTTP Request
- Data is always up-to-date without manual interaction
Search Engines - Document Retrieval

- Search Engine Query String:
  "Jaguar of 2000"

- Refine 'jaguar'
  CAR or ANIMAL or OPERATINGSYSTEM etc.
  Refine: CAR
Search Engines - Document Retrieval

- Search Engine Query String: "Jaguar of 2000"

- Refine '2000'
  ConstructionYear or technicalParameter?
  Refine: **ConstructionYear**

- Name: Jaguar E Type
- Category: Car
  - ...
  - ConstructionYear <Integer>
  - technicalParameter
    - CylinderCapacity <Integer>
    - Power <Integer>
  - ...

Knowledge Representation (Ontology, Linked Data)
Search Engines - Document Retrieval

- General Problems:
  - correct interpretation of query string
  - correct identification of entities
  - automatic disambiguation
  - usability
  - personalization
Search Engines - Fact Retrieval

- Query String:

  “Where can I fill up the tank at a considerable discount?”

Answer:
- Hohenfelden, xy-Str. 32 -> Super leaded, 1,99 €
- fuel-efficient route will be passed to navigation
- drive only at half throttle for saving fuel...
Search Engines - Exploratory Search

- dbpedia:Michael_Collins
- dbpedia-owl:mission
- dbpedia:Apollo_11
- dcterms:subject
- dbpedia-owl:mission
- dbpedia:Buzz_Aldrin
- dcterms:subject
- category:Apollo_program
- dbpedia:Apollo_13
- rdf:type
- dbpedia:Space_Shuttle_Challenger
- rdf:type
- yago:Space_accidents_and_incidents
Intelligent Agents in the Semantic Web

WWW

User

presentation service (e.g. Firefox)
retrieval service (e.g. Google)
WWW documents

Semantic Web

User

personal assistant
intelligent infrastructure services
WWW documents
3 Generations of Web Documents

1. Generation
- static web pages
  - HTML / CSS

2. Generation
- interactive web pages
  - JavaScript / Applets
  - Data Base Access
  - Template-based Generation

3. Generation
- virtual web pages
  - Netbots
  - Information Extraction
  - Presentation Planning

- adaptive web pages
  - User Models
  - Machine Learning
  - Online Layout