Topics of data engineering

Session 13

Masaomi Kimura

Today’s topic

- What is XML?
- Sementic Web

XML

XML is …

- eXtensible Markup Language
    - whose origin is SGML
- in a text format
    - XML uses tags as metadata of character strings
        - One can freely define the names of tags
        - Tags need to be closed if they are opened
        - It is possible to assign attributes to tags
    - XML is useful to exchange data, since textfiles can be read in any OS.

Is HTML a XML document? - No

- Both HTML and XML are markup languages using tags
  - Their origin is common, SGML
- A relationship between tags and content strings therein
  - Tags of HTML are for display, not giving meaning
  - Tags of XML are metadata to define contents' meaning
- Constraints for tags
  - In HTML, some tags, <P> and <BR>, are not required to be closed
  - In XML, all tags need to be closed

Elements and attributes

- Elements
  - A unit surrounded by a tag in a XML document
    - <name>Kimura</name>
  - Empty element
    - <name />   =   <name></name>
- Attributes
  - Additional information in elements
  - are included in start-tags
    - <name staff=“yes” >Kimura</name>

Structure of XML

XML declaration

```
<?xml version="1.0" encoding="Shift_JIS"?>
<!DOCTYPE book[
        <!ELEMENT book (bookname,author+)>
                <!ELEMENT bookname (#PCDATA)>
                <!ELEMENT author (name)>
                <!ELEMENT name (#PCDATA)>
                <!ATTLIST book format (paperback | hardback) "paperback">
]>
<book  format="hardback" >
        <bookname>XML for dummy</bookname>
        <author>
                <name>Kaori Takanashi</name>
        </author>
        <author>
                <name>Tatsuya Kimura</name>
        </author>
</book>
```

DTD

attributes

elements

XML declaration

- The declaration that the document is XML
  - necessary
  - <?xml version=“1.0”?>
- contains
  - version
    - version=“1.0”
  - encoding
    - encoding=“Shift_JIS”
    - if encoding is UTF-8, this is optional
  - standalone or not
    - standalone=“no” (default)
    - optional

DTD

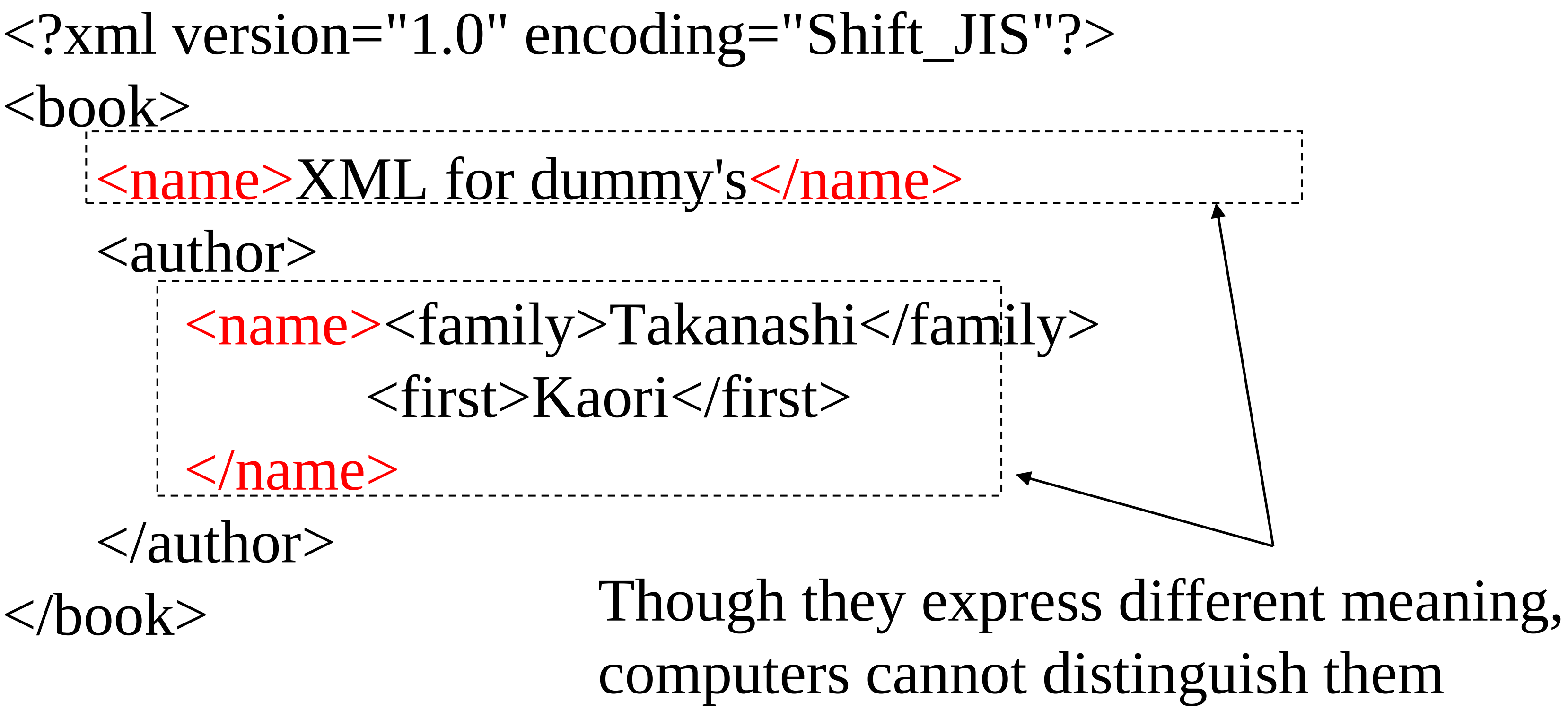
- □ defines a structure of XML

```
<!DOCTYPE root element [
    <!ELEMENT element （child elements)>
                ・・・
    <!ATTLIST element attribute value default>
                ・・・
]>
```

Namespaces

- Sets of names of elements and attributes
  - are necessary to avoid name conflicts
    - e.g. the cases if we want to use “name” tags to express the names of books and the names of authors
  - are expressed using a name prefix and are identified by URI

Without namespaces



```
<?xml version="1.0" encoding="Shift_JIS"?>
<book>
    <name>XML for dummy's</name>
    <author>
        <name><family>Takanashi</family>
                <first>Kaori</first>
        </name>
    </author>
</book>
```

Though they express different meaning, computers cannot distinguish them

With namespaces

```
<?xml version="1.0" encoding="Shift_JIS"?>
<b:book xmlns:b=“http://www.data.co.jp/book/”>
    <b:name>XML for dummy's</b:name>
    <a:author xmlns:a=“http://www.data.co.jp/author/”>
        <a:name><a:family>Takanashi</a:family>
                      <a:first>Kaori</a:first></a:name>
    </a:author>
</b:book>
```

XML schema

- Elements of XML should be defined in the form of XML
  - DTD has the same role, but its description is completely in different way.
  - DTD is too old to support namespaces
  - XML schemas have more degree of freedom to express iteration

Comparison of DTD and XML Schema

```
<!DOCTYPE book[
      <!ELEMENT book (bookname)>
            <!ELEMENT bookname (#PCDATA)>
]>
```

```
<?xml version="1.0"?>
<xsd:schema xmlns:xsd=http://www.w3.org/2001/XMLSchema
                    targetNamespace=http://www.data.co.jp/bookSchema/>
            <xsd:element name=“book">
                <xsd:complexType>
                        <xsd:element name=“bookname" type="xsd:string"/>
               </xsd:complexType>
           </xsd:element>
</xsd:schema>
```

Application of XML

SEMANTIC WEB

Tim Berners-Lee

- a British computer scientist and a father of WWW
  - He built a prototype system of WWW, ENQUIRE at CERN in 1980.
  - The director of W3C
- His idea of WWW is not only the current linked documents (HTML) but also linked data (Semantic Web).

https://www.ted.com/talks/tim_berners_lee_on_the_next_web

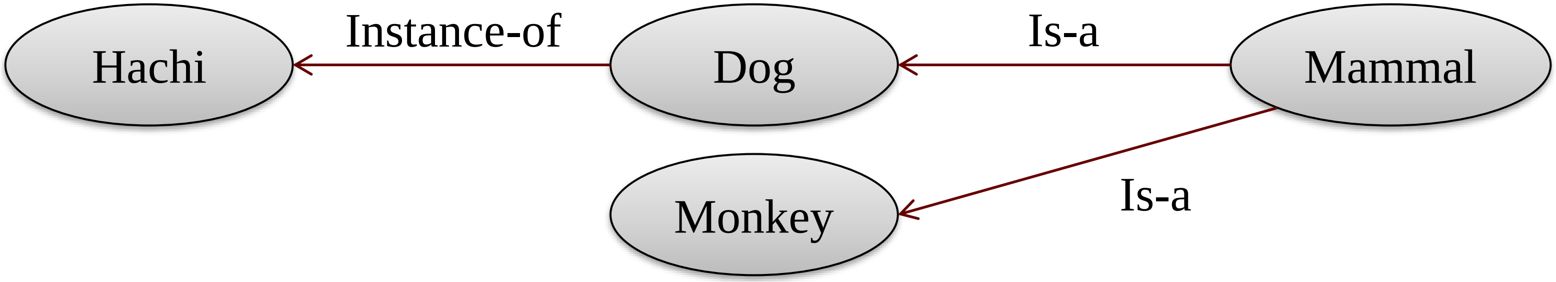
http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide

Semantic Web

- Semantics=a process to transfer meanings of things and concepts to support users’ action
- A complement of WWW to support computers’ reasoning/deduction based on ontology technique
- Its standard language is OWL
  - Web Ontology Language
    - Based on RDF

Ontology

- A theory of essence or existence
  - is originally a term in philosophy
  - defines a relationship between words
    - Is-a
      - A class is a subclass of another
    - Instance-of
      - A concept is an example of a class



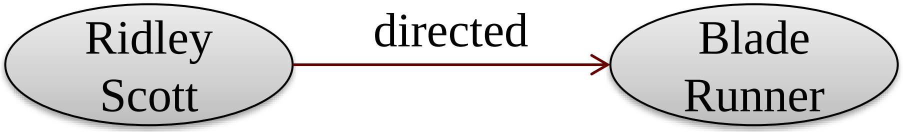
Ontology (in information science)

- Is a explicit and systematic specification of a shared concept
    - concept＝an abstract model in a real world
    - explicit＝the types of concepts and their constraints are defined
    - systematic＝understandable for computers
- CYC
    - http://www.cyc.com

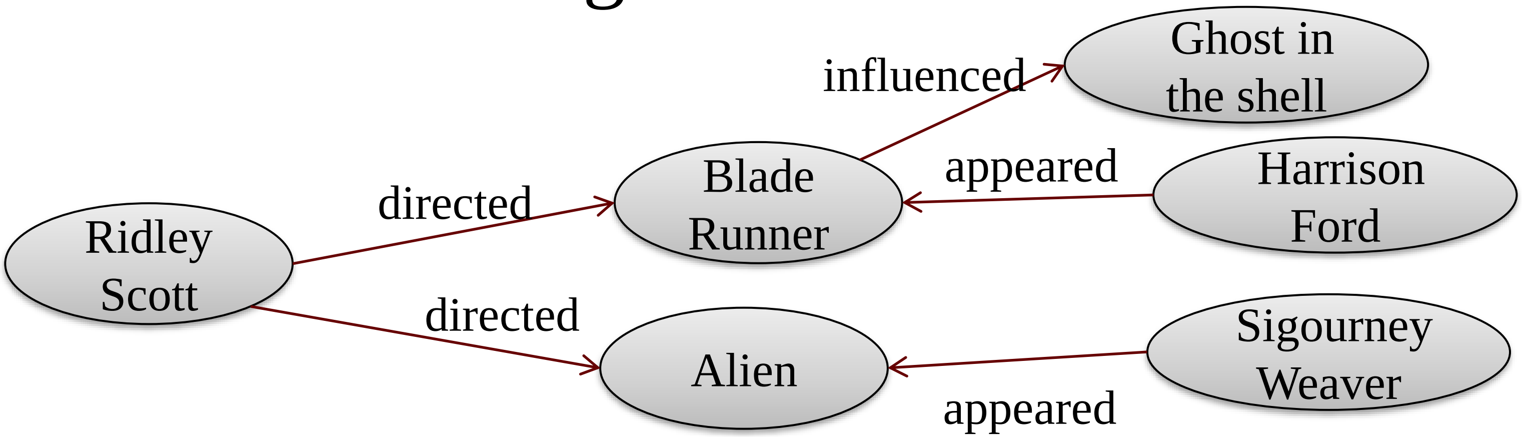
https://www.youtube.com/watch?v=iYhJ7Mf2Oxs

Triple

- Ridley Scott directed Blade Runner



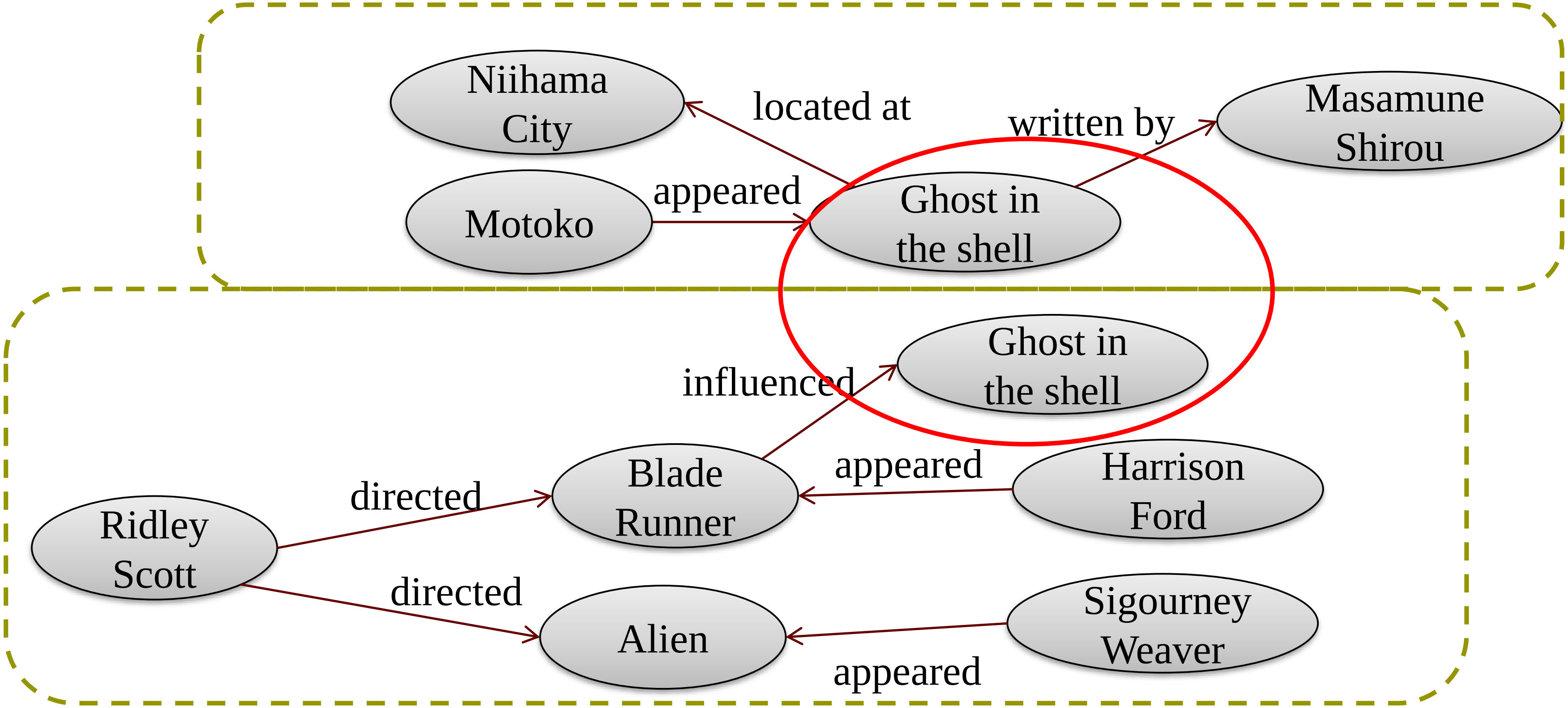
- There are many kinds of relationships, which needs flexible management



https://www.youtube.com/watch?v=p2MEaROKjaE

Combining of graphs

- Unique IDs (=URIs) strongly help combining the semantic graphs seamlessly

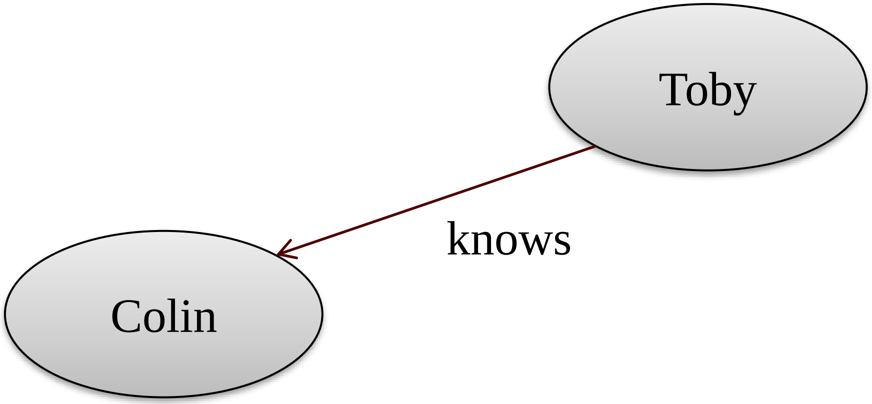


RDF

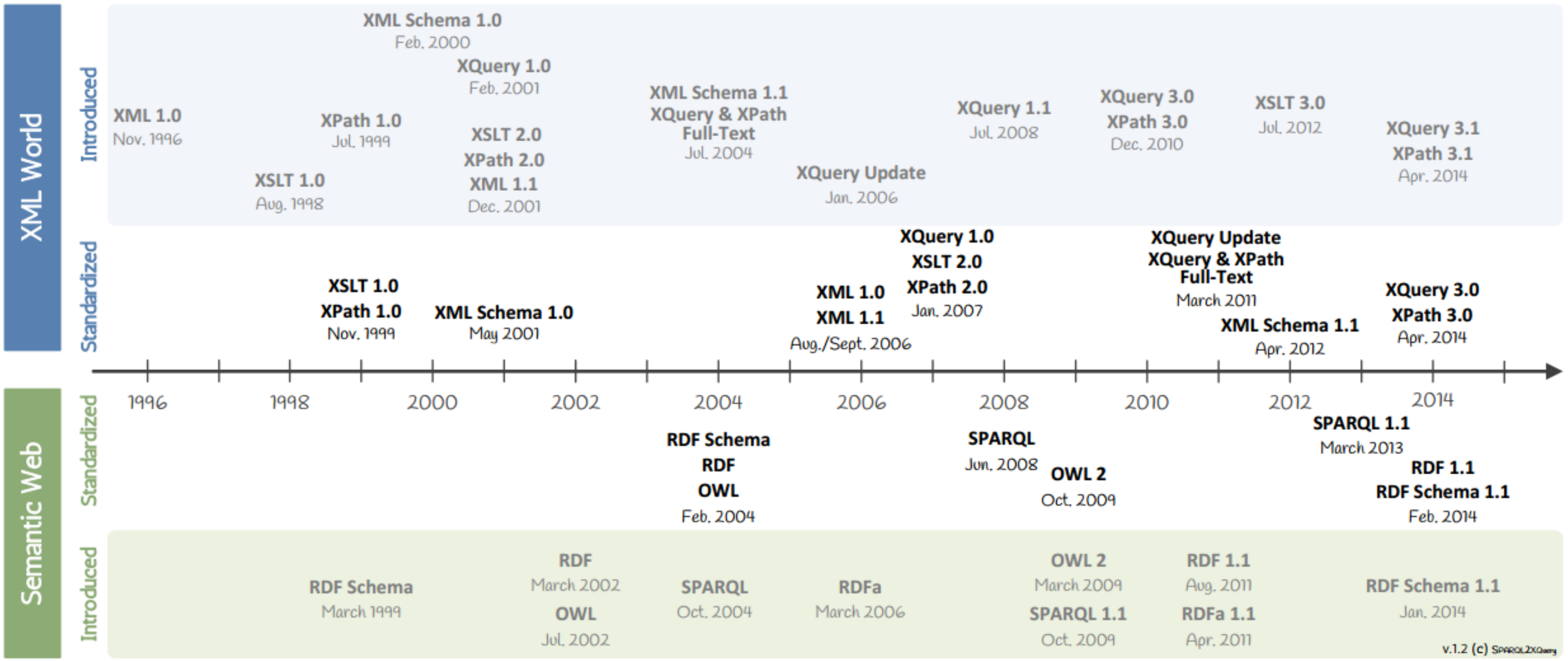
- Resource Description Framework
  - A standard of W3C to describe triples
  - Developped in 1990’s
- As a data model
  - URI is a key to identify a unique resource (thing/concept)
  - Describes triplets

An example of RDF documents

```
<rdf:RDF xmlns:rdf=“http://www.w3.org/1999/02/22-rdf-syntax-ns#
              xmlns:foaf=“http://xmlns.com/foaf/0.1/”>
   <rdf:Description rdf:about=“http://kiwitobes.com/toby.rdf#ts”>
      <foaf:knows>
          <rdf:Description rdf:About=http://semprog.com/people/colin>
              <foaf:name>Colin Evans</foaf:name>
           </rdf:Description>
      </foaf:knows>
   </rdf:Description>
</rdf:RDF>
```



Time line of RDF



XML World

Introduced

XML 1.0
Nov. 1996

XSLT 1.0
Aug. 1998

XPath 1.0
Jul. 1999

XML Schema 1.0
Feb. 2000

XQuery 1.0
Feb. 2001

XSLT 2.0
XPath 2.0
XML 1.1
Dec. 2001

XML Schema 1.1
XQuery & XPath
Full-Text
Jul. 2004

XQuery Update
Jan. 2006

XQuery 1.1
Jul. 2008

XQuery 3.0
XPath 3.0
Dec. 2010

XSLT 3.0
Jul. 2012

XQuery 3.1
XPath 3.1
Apr. 2014

Standardized

XSLT 1.0
XPath 1.0
Nov. 1999

XML Schema 1.0
May 2001

XML 1.0
XML 1.1
Aug./Sept. 2006

XQuery 1.0
XSLT 2.0
XPath 2.0
Jan. 2007

XQuery Update
XQuery & XPath
Full-Text
March 2011

XML Schema 1.1
Apr. 2012

XQuery 3.0
XPath 3.0
Apr. 2014

1996 1998 2000 2002 2004 2006 2008 2010 2012 2014

Semantic Web

Standardized

RDF Schema
RDF
OWL
Feb. 2004

SPARQL
Jun. 2008

OWL 2
Oct. 2009

SPARQL 1.1
March 2013

RDF 1.1
RDF Schema 1.1
Feb. 2014

Introduced

RDF Schema
March 1999

RDF
March 2002
OWL
Jul. 2002

SPARQL
Oct. 2004

RDFa
March 2006

OWL 2
March 2009
SPARQL 1.1
Oct. 2009

RDF 1.1
Aug. 2011
RDFa 1.1
Apr. 2011

RDF Schema 1.1
Jan. 2014

v.1.2 (c) SPARQL2XQuery

http://www.dblab.ntua.gr/~bikakis/XML%20and%20Semantic%20Web%20W3C%20Standards%20Timeline-History.pdf