Semistructured Data

Structured Data

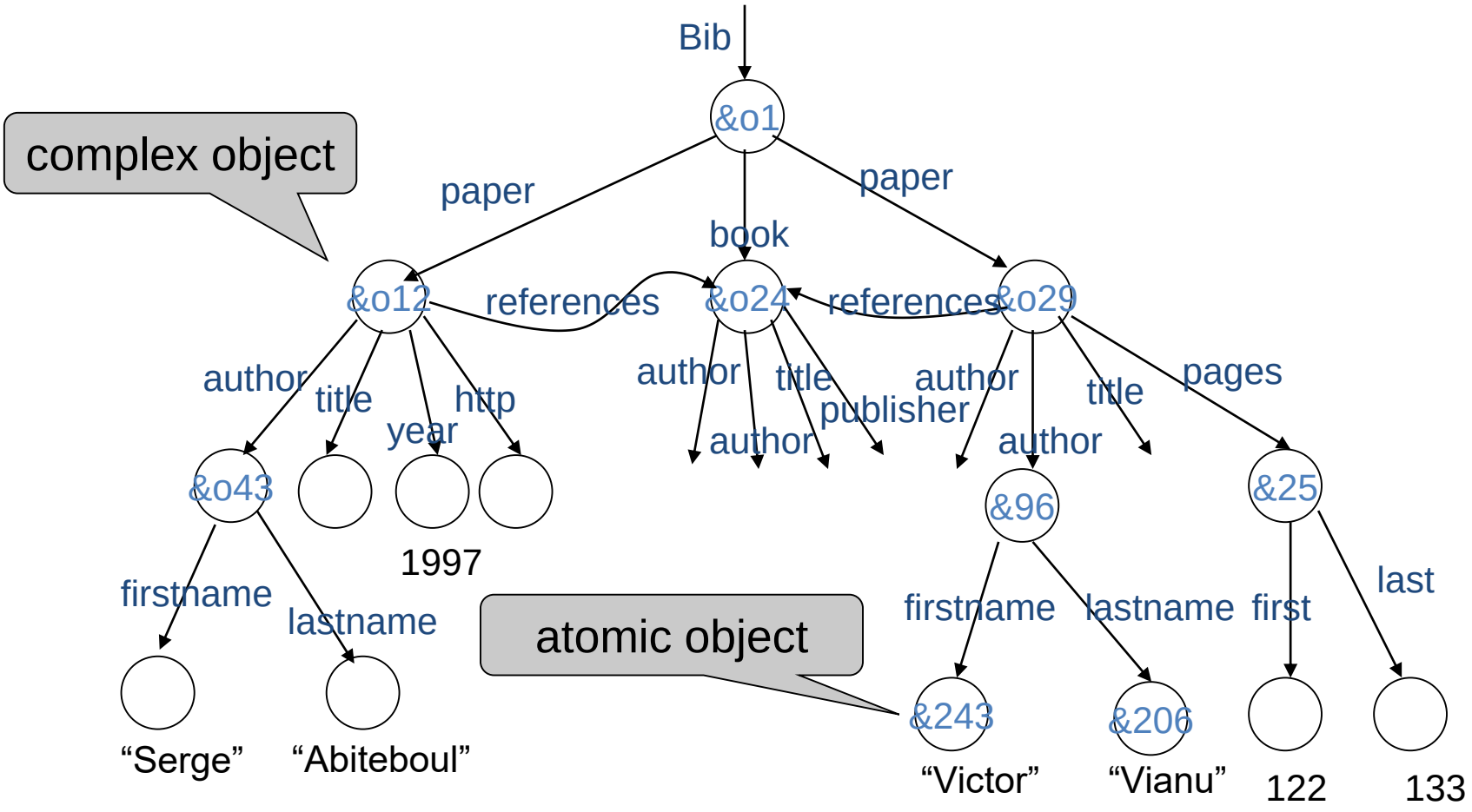
ID	Last Name	First Name	Title	Birth Date	Hire Date	City	Region
1	Davolio	Nancy	Ms.	08-dic-1968	01-mag-1992	Seattle	WA
2	Fuller	Andrew	Dr.	19-feb-1952	14-ago-1992	Tacoma	WA
3	Leverling	Janet	Ms.	30-ago-1963	01-apr-1992	Kirkland	WA
4	Peacock	Margaret	Mrs.	19-set-1958	03-mag-1993	Redmond	WA
5	Buchanan	Steven	Mr.	04-mar-1955	17-ott-1993	London	
6	Suyama	Michael	Mr.	02-lug-1963	17-ott-1993	London	
7	King	Robert	Mr.	29-mag-1960	02-gen-1994	London	

Order ID	Customer	Emp ID	Order Date	Required Date	Shipped Date
10248	Wilman Kala	1	04-lug-1996	01-ago-1996	16-lug-1996
10249	Tradição Hiperm.	6	05-lug-1996	16-ago-1996	10-lug-1996
10250	Hanari Carnes	3	08-lug-1996	05-ago-1996	12-lug-1996
10251	Victuailles en stock	3	08-lug-1996	05-ago-1996	15-lug-1996
10252	Suprêmes délices	2	09-lug-1996	06-ago-1996	11-lug-1996
10253	Hanari Carnes	3	10-lug-1996	24-lug-1996	
10254	Chop-suey Chinese	2	11-lug-1996	08-ago-1996	23-lug-1996

Unstructured data

- ***Sample databases included with Access***
  - Microsoft Access provides sample databases that you can use while you're learning Access.
  - **Northwind Traders sample database**
  - The Northwind database and Access project (available from the **Sample Databases** command on the **Help** menu) contains the sales data for a fictitious company called Northwind Traders, which imports and exports specialty foods from around the world. By viewing the **database objects** included in the Northwind database. …

Semistructured Data



Sources of SSD

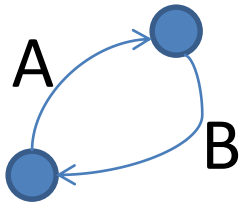
- Data integration
- Scientific data
- Documents
- WWW

SSD data models

- A data model:
  - The information behind the syntax (by a denotation or through equality)
  - The operators
- Some alternatives:
  - RDF: graphs with node equality – triples of URIs
  - OEM: graphs modulo bisimulation
  - XML: ordered trees with “node identity”, parent pointer, fusion of blanks in text nodes
  - JSON: unordered trees with ordered arrays

Graphs

- Graph with labelled edges:
  - (N,E,f) con $E \subseteq (N \times N)$ ed $f:E \rightarrow \Phi$
- Graph equality: node isomorphism
  - ({1,2}; {(1,2),(2,1)}; {(1,2)→A; (2,1)→B})
  - ({a,b}; {(a,b),(b,a)}; {(a,b)→A; (b,a)→B})
  - ({a,b}; {(a,b),(b,a)}; {(a,b)→B; (b,a)→A})



Isomorphism

- Isomorphism between (E,N,Φ) and (E',N',Φ'): bijection σ from N to N' such that:
  - $(n,m) \in E$ iff $(\sigma(n), \sigma(m)) \in E'$
  - $(n,m) \in E \Rightarrow \Phi(n,m) = \Phi'(\sigma(n), \sigma(m))$
- We say that nodes are 'anonymous', have 'no identity'

RDF

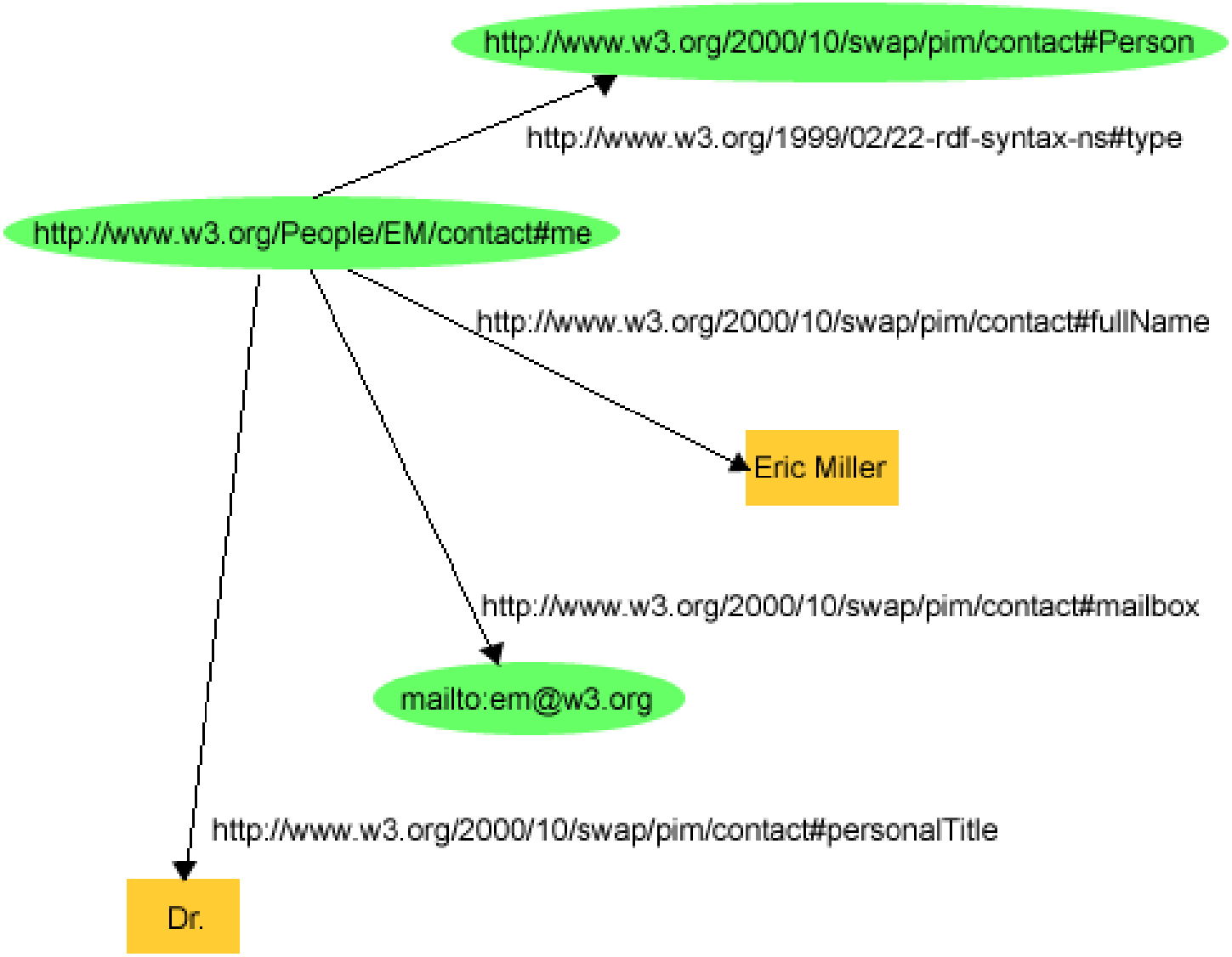
- A graph is a set of triples:

<http://www.w3.org/People/EM/contact#me>
<http//www.w3.org/…/22-rdf-syntax-ns#type>
<http://www.w3.org/swap/pim/contact#Person>

<http://www.w3.org/People/EM/contact#me>
<http://www.w3.org/swap/pim/contact#fullName>
"Henry Miller"

- Nodes are *not* anonymous

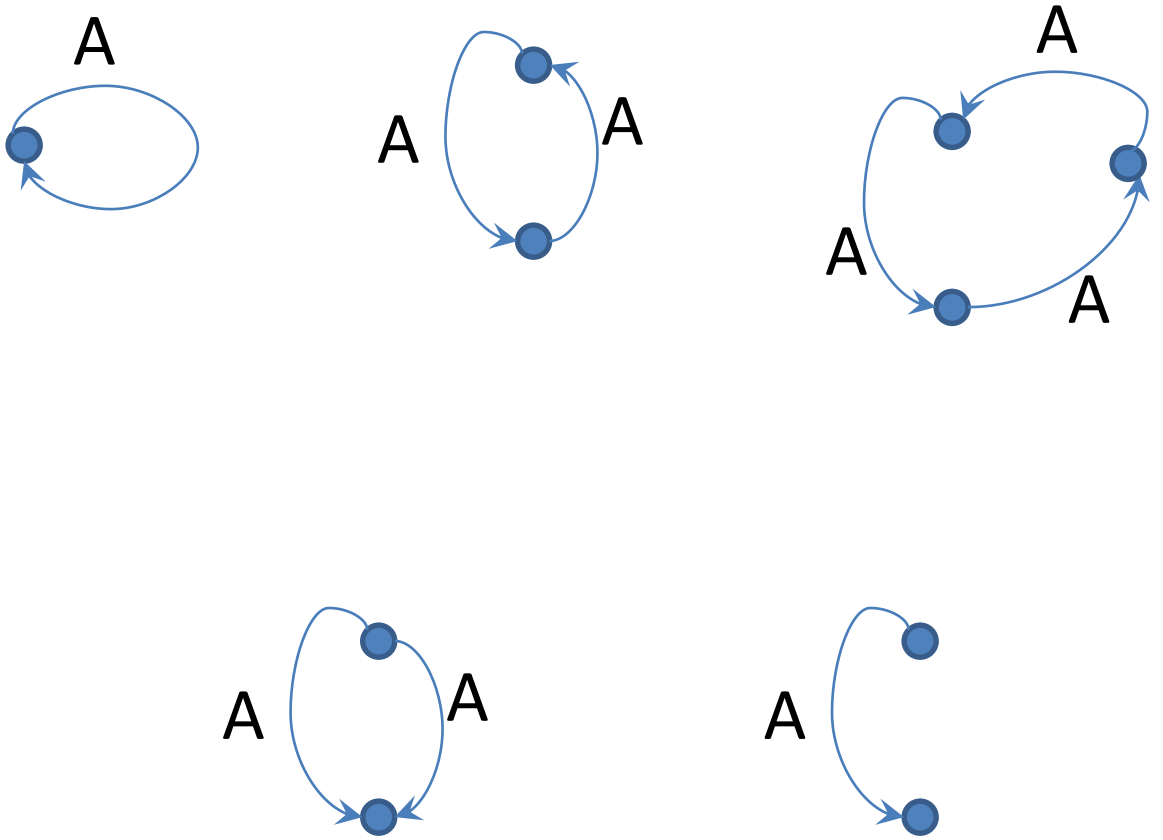
RDF graph


http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

OEM

- Graphs with labelled edges and anonymous nodes (leaves are labelled with atomic values)
- Bisimulation: set equality generalized to graphs
  - {a: v, b: w} = {b: w, a: v}
  - {a: v, a: v, b: w} = {a: v, b: w, b: w}
- Formally: exists $R \subseteq G \times G'$ such that:
  - n R m and n,l,n' in G $\Rightarrow$ exists m,l,m' in G' with n' R m' and vice versa
  - n R m and n leaf in G $\Leftrightarrow$ m leaf in G'
- The idea: paths are observable, nodes are not, multiplicity of paths is not

Bisimulation



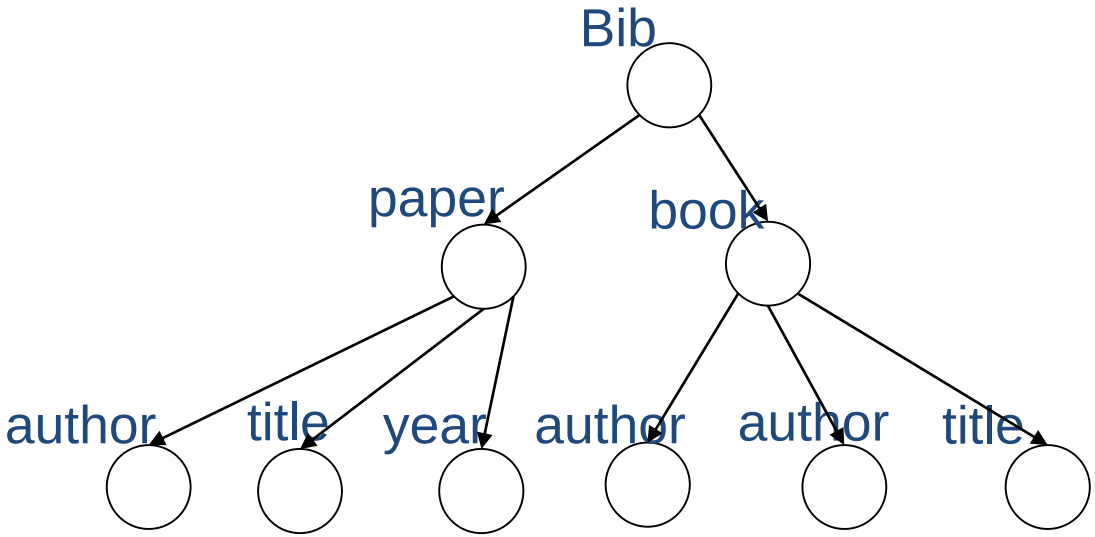
XDM (XML Data Model)

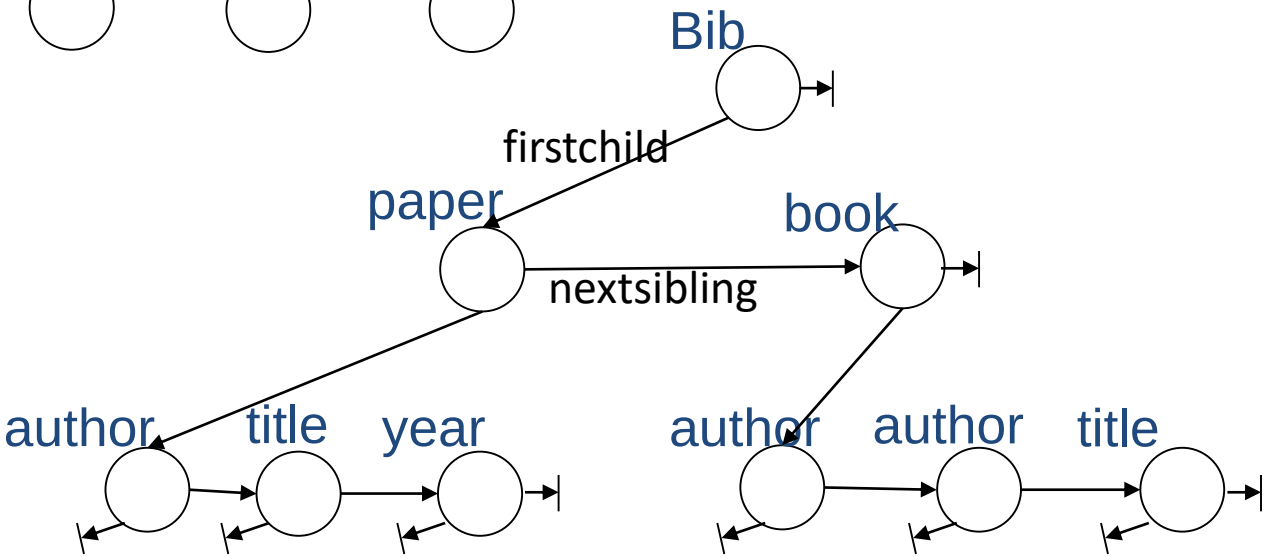
- Node-labelled ordered trees:
  - {a: v, b: w} ≠ {b: w, a: v}
  - {a: v, a: v, b: w} ≠ {a: v, b: w, b: w}

Ordered trees and binary trees

- Binary trees:
  - bt ::= _ | label[bt,bt]
  - Ad es.: a[b[_,_],_] ≠ a[_,b[_,_]]
- Ordered forests are isomorphic to binary trees

Ordered forests and binary trees





XDM: further details

- Every node has an identity, that can be compared to other nodes identity
- Who is the parent of a[b,c,d]?
- In XDM, you can navigate from a node to its parent
- An XDM tree is actually a pair
  - A whole tree (the ‘document’)
  - A pointer inside that tree (the ‘current node’)

XDM: further details

- 7 types of nodes: elements, attributes (unordered), text nodes, …

JSON

```
<menu id="file">
        <popup>
                <menuitem value="New" onclick="CreateNew()" />
                <menuitem value="Close" onclick="CloseDoc()" />
        </popup>
</menu>
```

```
{"menu": {
        "id": "file",
        "popup": {
                "menuitem": [
                        {"value": "New", "onclick": "CreateNew()"},
                        {"value": "Close", "onclick": "CloseDoc()"}
                ]
        }
}}
```

JSON

- http://json.org/

object ::= **{}** | **{** (string **:** value**,**)* string : value**}**
array ::= **[]** | **[** (value **,**)* value **]**
value ::= object
| array
| string
| number
| **true** | **false** | **null**
string ::= **"** char * **"**
number ::= ….

- Objects are unordered
- The names in an object *should* be unique

The object model

- Similar to OEM
- However:
  - A priori schema, every object belong to a class, the class specifies the outward labels
  - Classes have methods

Other models

- TQL
  - Moltiplicity is observable, order is not:
    - {a: v, b: w} = {b: w, a: v}
    - {a: v, a: v, b: w} ≠ {a: v, b: w, b: w}
  - Si possono interpretare come alberi con archi etichettati oppure foreste con nodi etichettati
- Compositional graphs:
  - (x,a,y)|(x,b,z)
  - (νx.(x,a,y)|(x,b,z))
  - (νx.(x,a,y)|(x,b,z)) | (νx.(x,a,y)|(x,b,z))

Sources

- S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web*, Morgan Kaufman, 2000, Chapters 1-2
- Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, et al., Web Data Management, Cambridge University Press, 2011, Chapter 1, http://webdam.inria.fr/Jorge/
- RDF: http://www.w3.org/TR/rdf-primer/
- Json.org