UNDERSTANDING DATA

By Kingsley Idehen
Founder & CEO, OpenLink Software

OPENLINK SOFTWARE
Making Technology Work For You®

Presentation Goals

Deconstruct Data

Understand Data Representation

Understand Data Access

Understand Data Integration

SITUATION ANALYSIS

EVERY DAY WE HEAR

DATA IS
BIG

DATA IS
OPEN

DATA IS
LINKED

WE ALMOST **NEVER** HEAR ABOUT

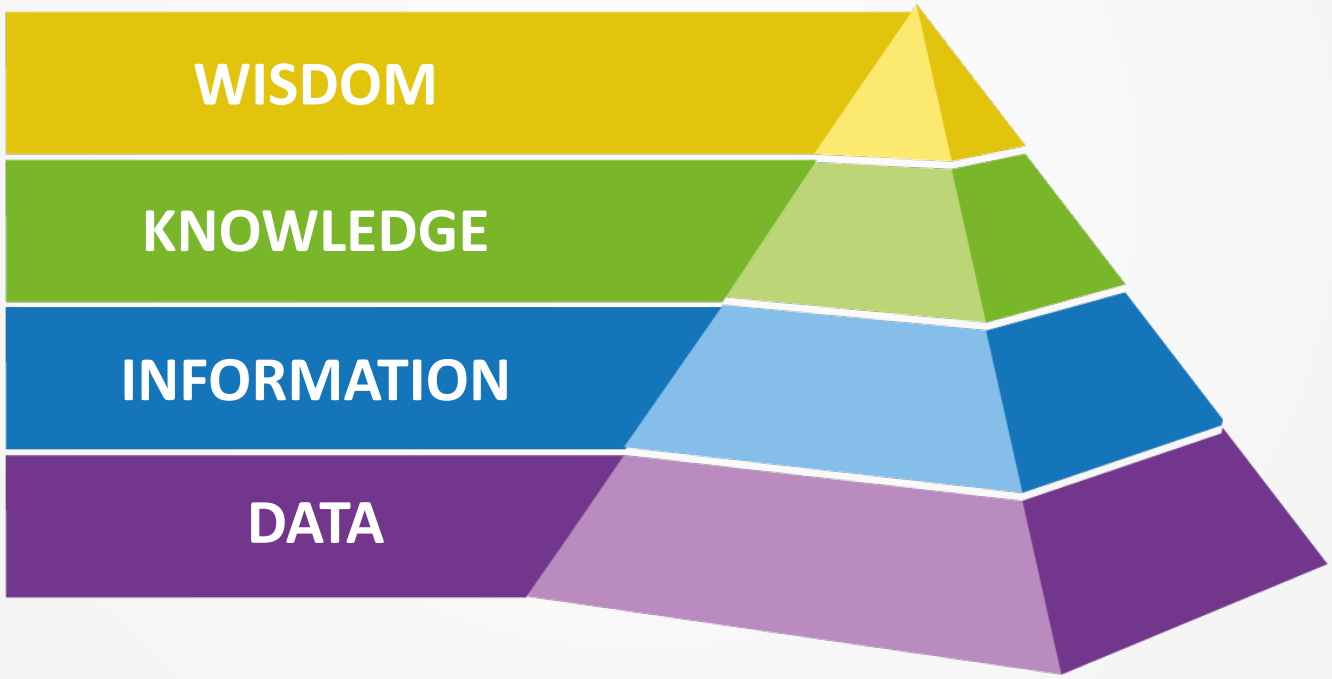
WHAT DATA
ACTUALLY IS

HOW DATA IS
REPRESENTED

HOW DATA IS
ACCESSED,
SHARED,
& INTEGRATED

Why is Data Important?

Data is the basis of
Information, Knowledge, and Wisdom.

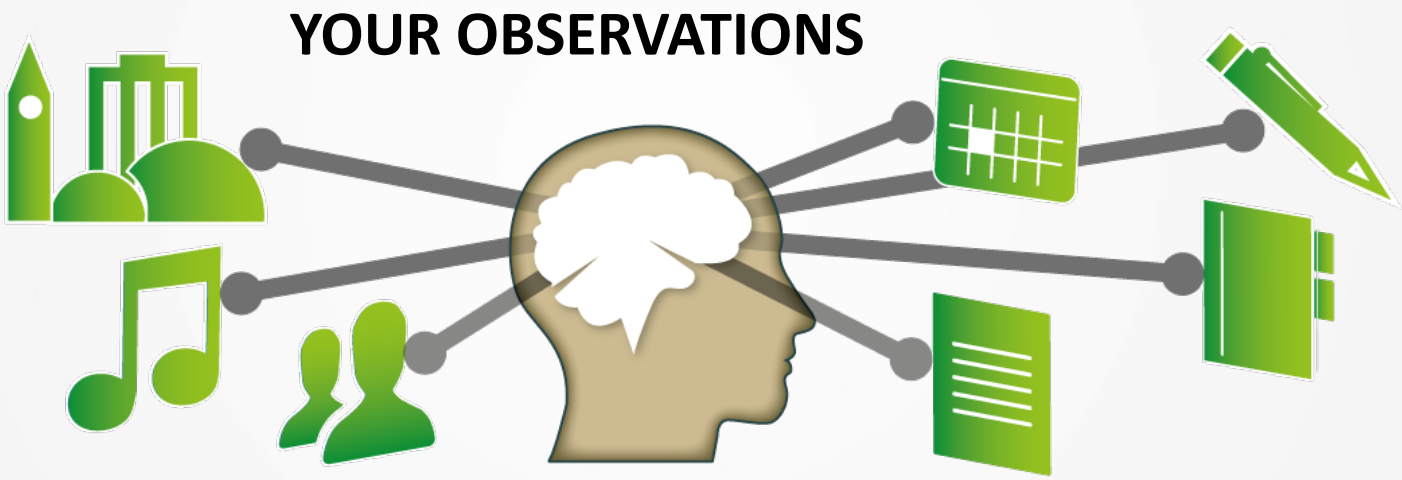


What is Data?

Data is how we express Observation in reusable form.

What is Observation?

Observation is the Perception of Relationships between Entities.



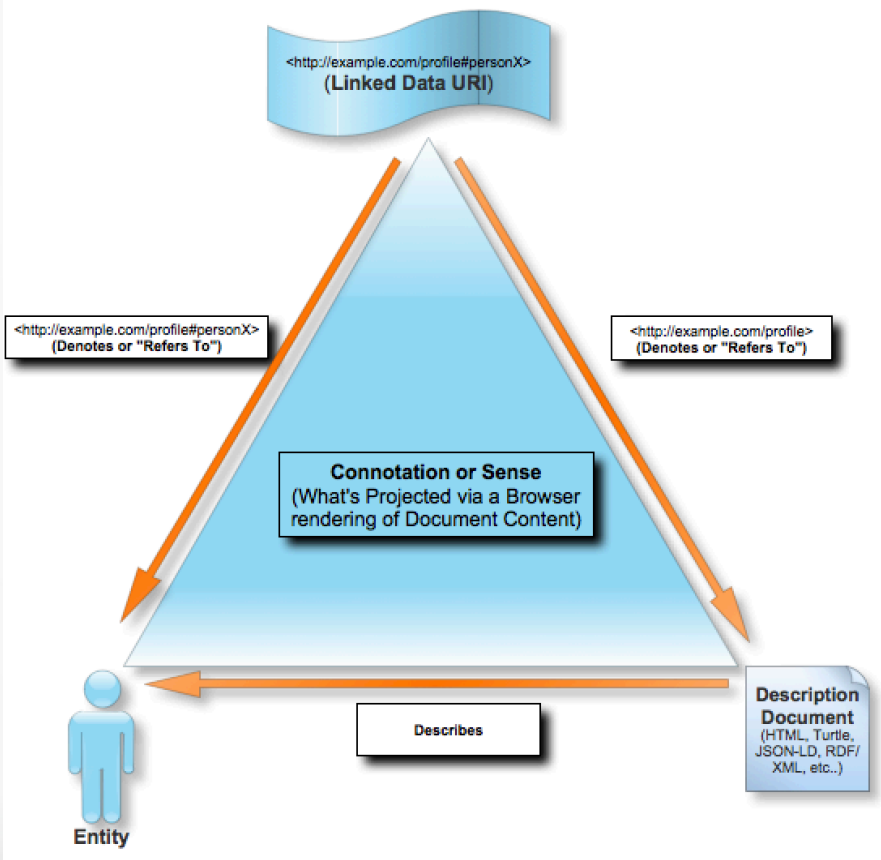
PEOPLE, PLACES, MUSIC, DOCUMENTS, CALENDARS, DIARIES, ADDRESS BOOKS & MORE...

What is an Entity?

An Entity is a Distinctly Identifiable Thing

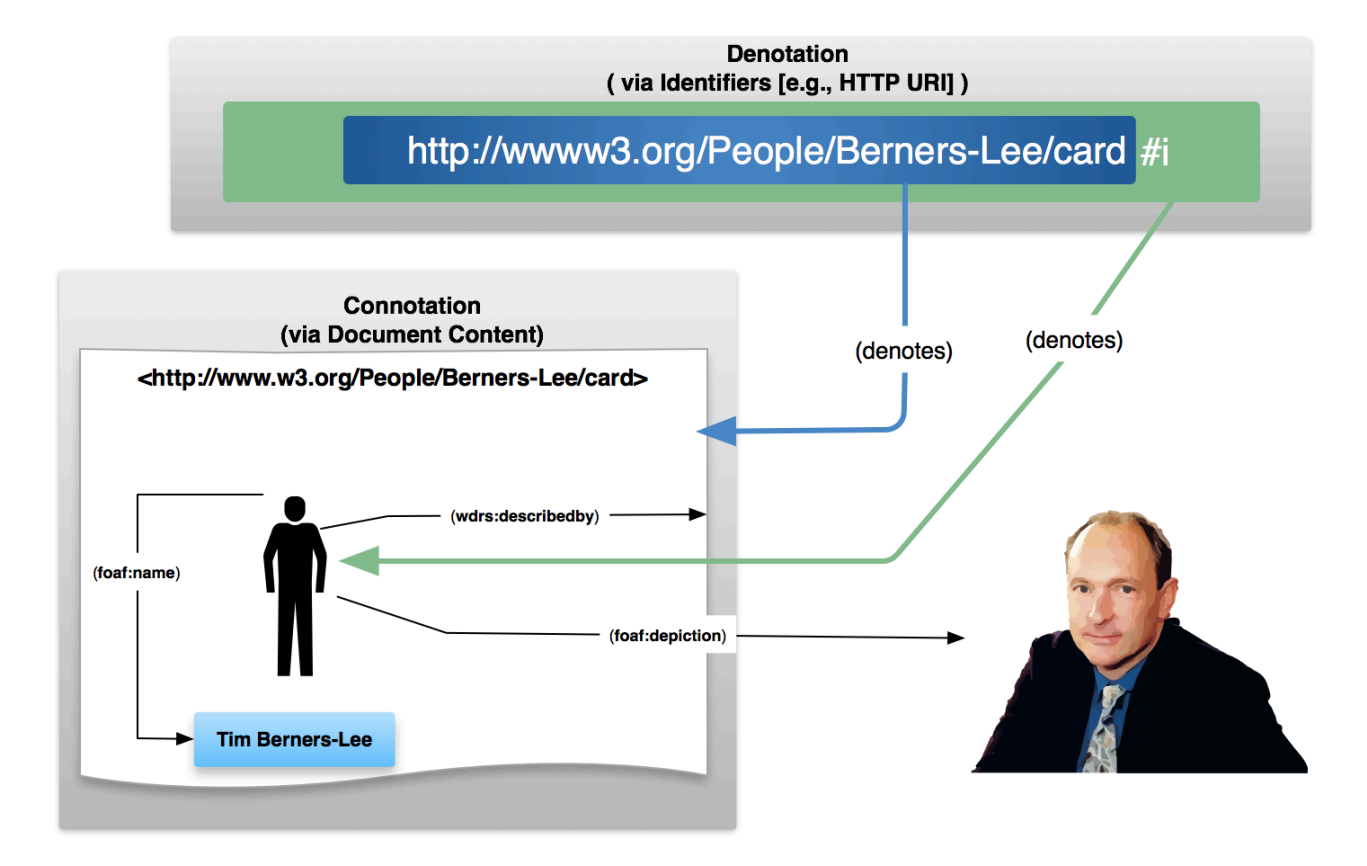
How is an Entity Identified (Named) ?

An Entity is Identified (or named) through the combined effects of Identifier based denotation (signification) and document content based connotation (description).



How Does Entity Identification Work?

Through interpretation that's driven by sign [denotation] -> description [connotation] based indirection.



How is an Entity Denoted?

An Entity is Denoted (Signified) through the use of an Identifier.

What is an Identifier?

An Identifier is a Sign (or Token) that Signifies (Denotes, or

"Stands For") an Entity

Identifier Types?

Quoted Literals such as:

“Kingsley Idehen” or ‘Kingsley Idehen’

Absolute References:

<http://kingsley.idehen.net/dataspace/person/kidehen#this>

Relative References:

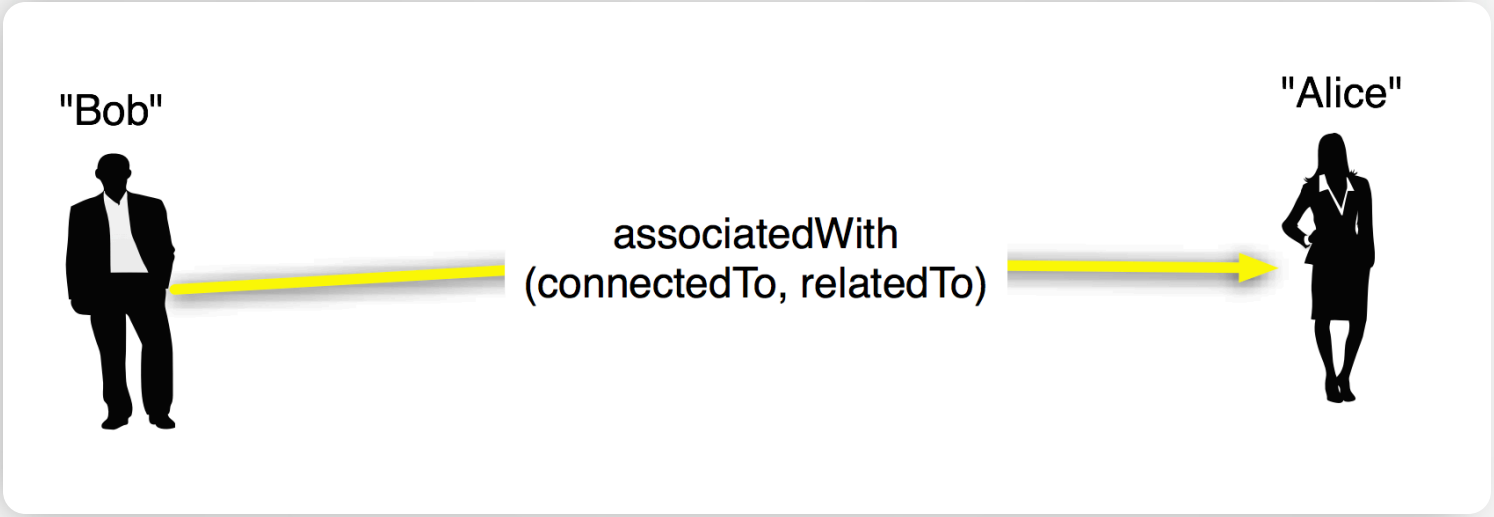
<#KingsleyIdehen>

How is an Entity Described?

Through entity relationships that are represented in reusable form via document content (sentences and statements).

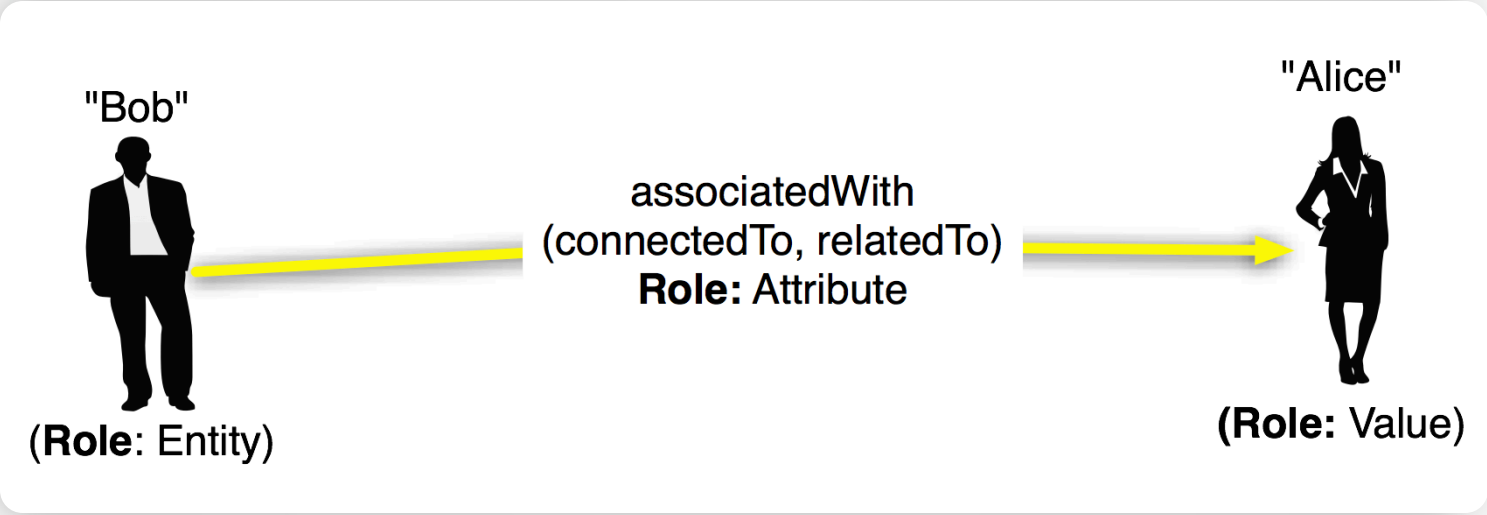
What is a Relationship?

A Relationship is an Association between two or more Entities, where each has a specific Role.



What is a Relationship Role?

A Relationship Role is a
Function performed by an Entity
in a Relationship

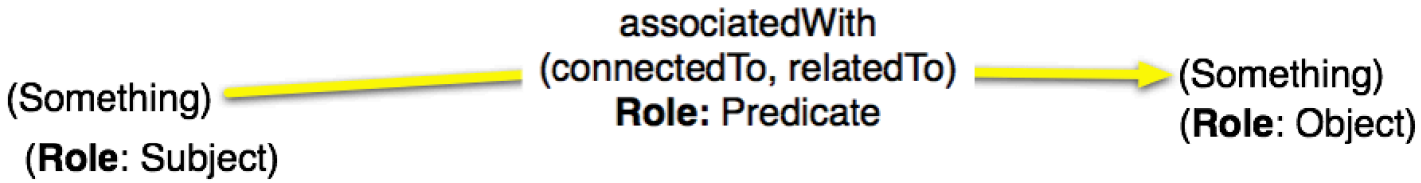


Relationship Role Types?

- **Entity Attribute Value EAV**
  - ✓ **Entity** -- observation focal point
  - ✓ **Attribute** -- observation attribute name (relationship type determinant)
  - ✓ **Value** -- observation attribute value
- **RDF (WC3's Resource Description Framework)**
  - ✓ **Subject** -- observation focal point
  - ✓ **Predicate** -- observation attribute name (relationship type determinant)
  - ✓ **Object** -- observation attribute value

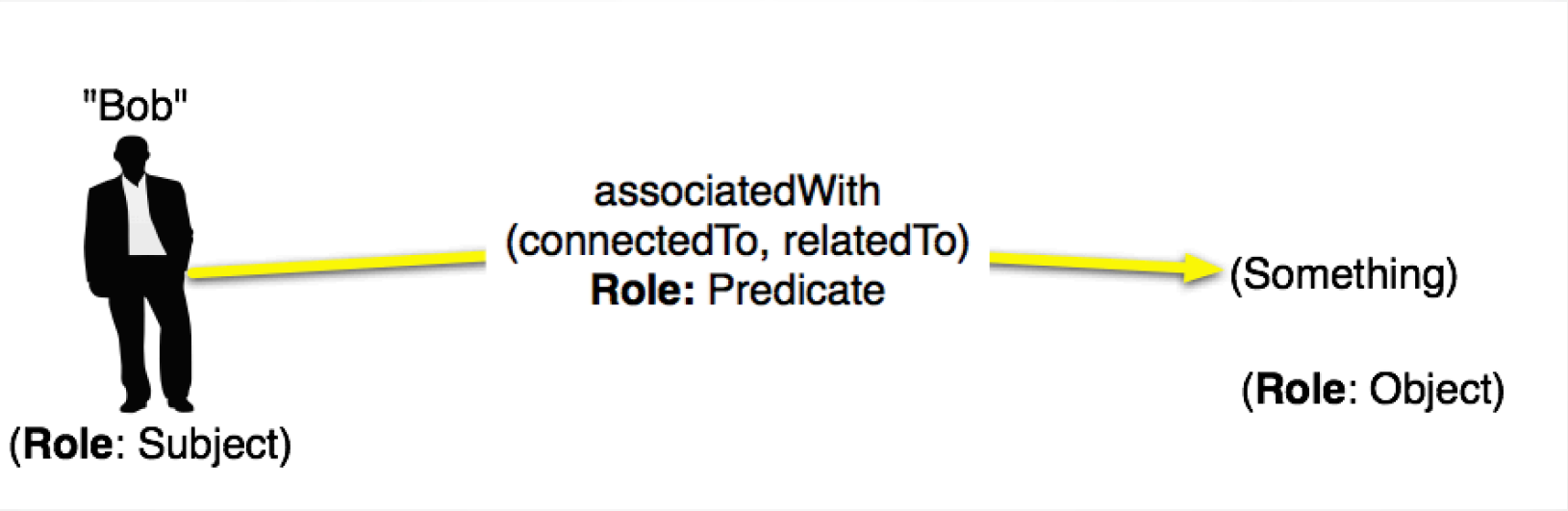
Relationship Role: Predicate

The Relationship Predicate is the Connector that associates an observation focal point (Subject) with something, in the form of an observation value (Object).



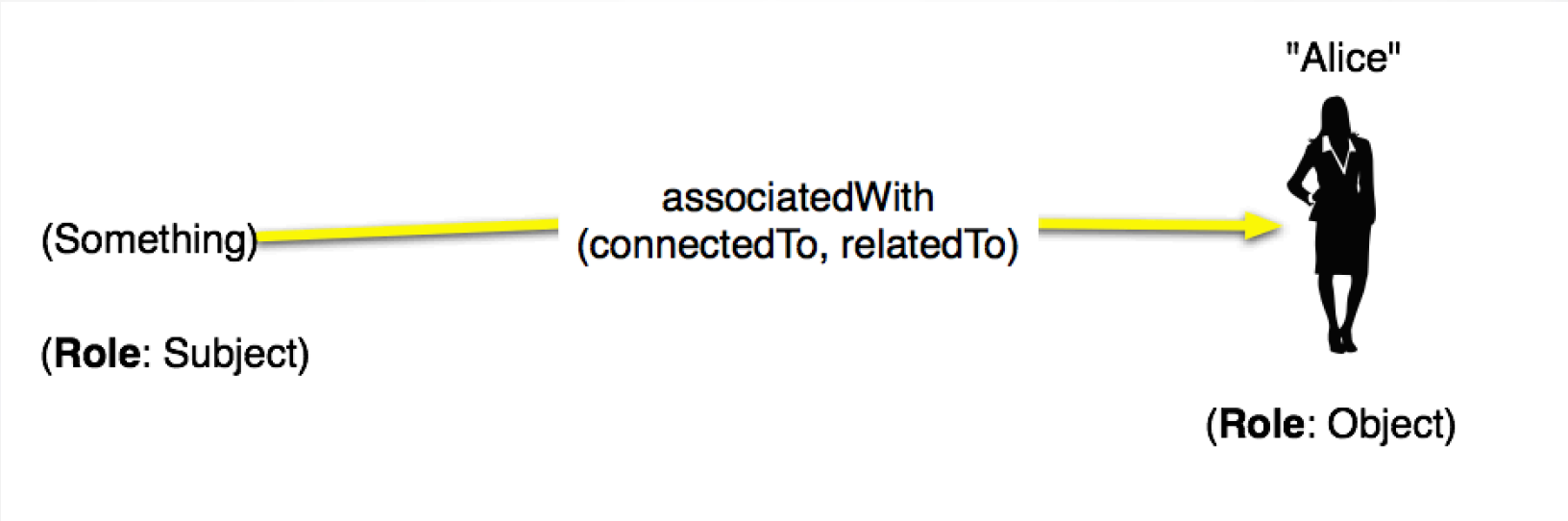
Relationship Role: Subject

Actual Entity being Observed



Relationship Role: Object

Value associated
with an observation focal point (Subject)
via a Relationship Predicate.



Types of Values?

- Untyped Literals (Strings)
- Typed Literals
  - ✓ Numbers
  - ✓ Dates
  - ✓ Booleans
  - ✓ Etc.
- References (Local and Global Hyperlinks)

How are Relationships Expressed?

Relationships are Expressed using a Language, i.e., a system of signs [for denotation], syntax [arrangement of signs to form sentences], and entity relation semantics [meaning of relationship roles] for encoding and decoding information.

Example:

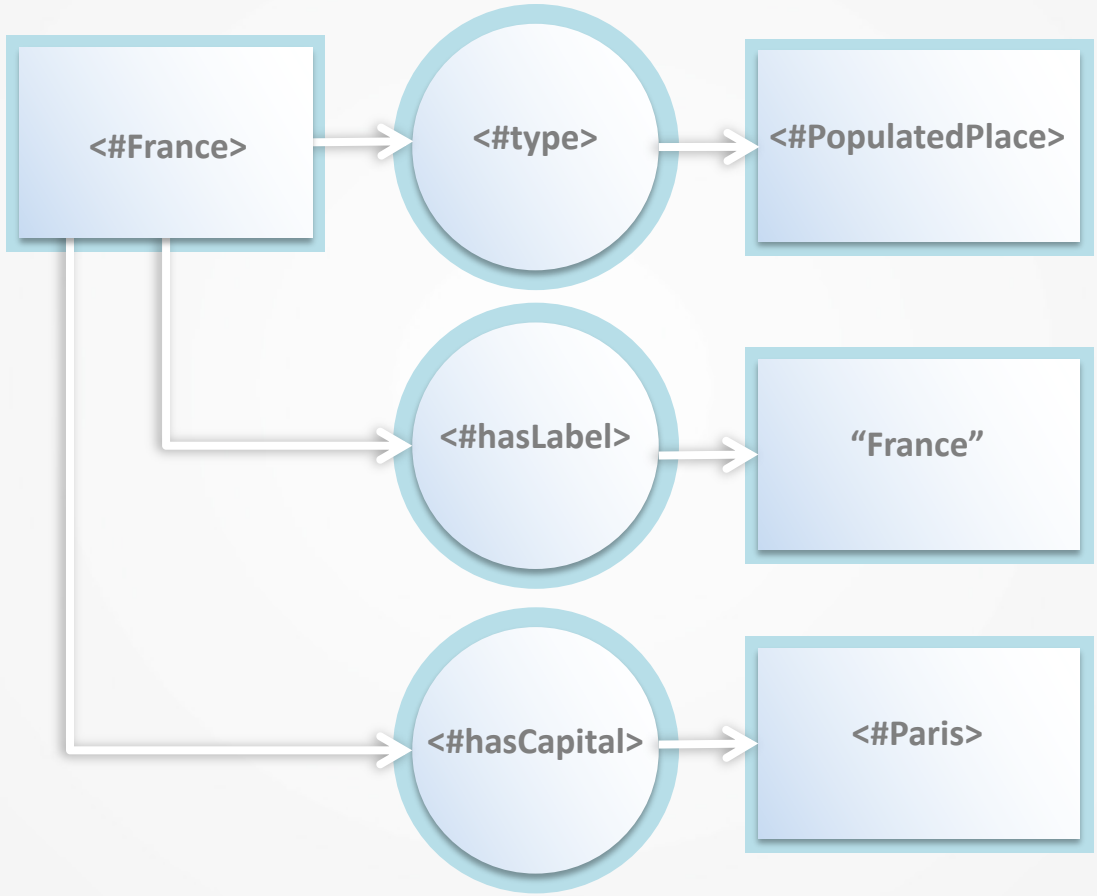
Subject, Predicate, Object –
Used by W3C's Resource Description Framework (RDF) and Natural Language.

How Are Entity Relationships Represented ?

Entity Relationships are Represented using notations associated with a specific language. Examples include:

- Entity Relationship Model (Network / Graph) Diagrams.
- Tables (CSV files, Spreadsheets, and SQL Relational Database Management Systems).
- RDF-Turtle, JSON-LD, RDF/XML, HTML+Microdata, HTML+RDFa etc..

Entity Relationship Diagram



Turtle Notation Based Entity Relationship Statements

```
<#France> <#Type> <#PopulatedPlace> .

<#France> <#hasLabel> ”France” .

<#France> <#hasCapital> <#Paris> .

<#Paris> <#Type> <#PopulatedPlace> .

<#Paris> <#hasLabel> ”Paris” .

<#PopulatedPlace> <#Type> <#Place> .
```

Entity Relationship Tables

Delimiter: e.g., Comma
Identifier Quote Character: Double-quotes
Relation Header Row: Entity,Attribute,Value
Relation Body

Example:
“Entity”, “Attribute” “Value”
“France”, “Type” “PopulatedPlace”
“France” , “hasLabel” “France”
“France” , “hasCapital” “Paris”

Statement Representation: Spreadsheet Tables

Entity (Subject)	Attribute (Predicate)	Value (Object)
#France	#Type	#PopulatedPlace
#France	#hasLabel	“France”
#France	#hasCapital	#Paris
#Paris	#Type	#PopulatedPlace
#Paris	#hasLabel	“Paris”
#PopulatedPlace	#Type	#Place

How are Statements Persisted & Transmitted?

- **Persistence:**
  - ✓ To paper based documents
  - ✓ To digital realm documents

    (e.g., operating system files, web pages, etc.)
- **Transmission:**
  - ✓ Text oriented serialization formats
  - ✓ Binary serialization formats

Understanding Data (Recap)

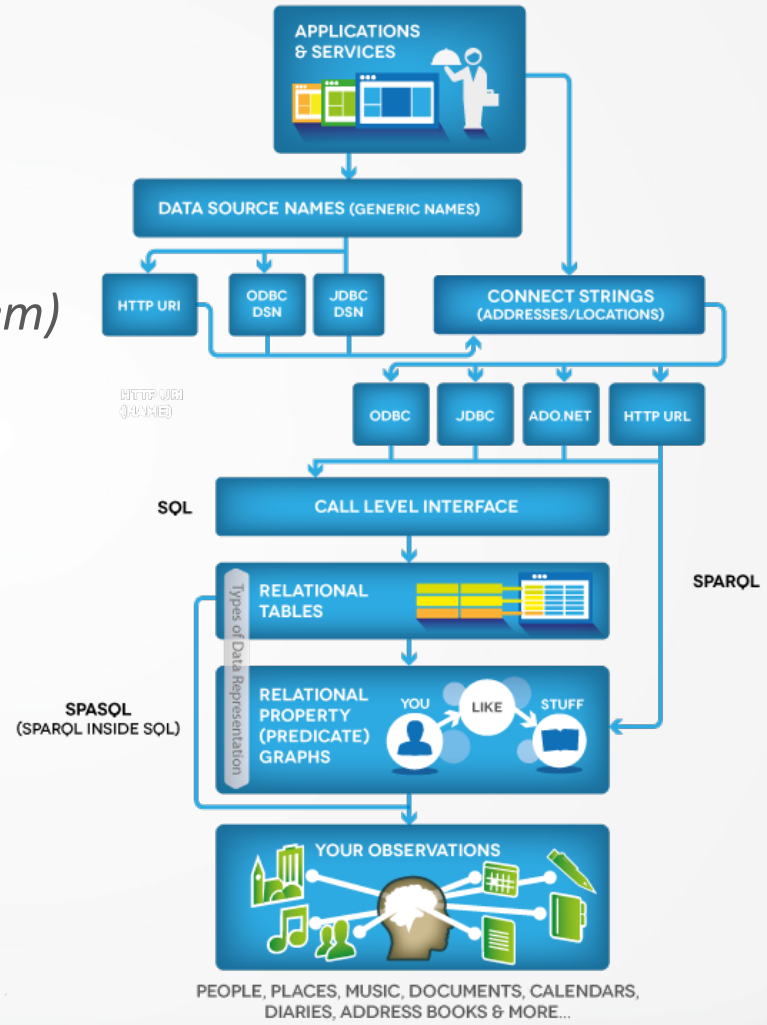
- The term **“Data”** refers to observation expressed in reusable form.
- The term **“Observation”** refers to our perception of Entity Relationships.
- **Entity Relationships** are expressed using a language.
- **Statements** are represented using a variety of notations; persisted to paper or digital documents; and transmissible using a variety of serialization formats.

DATA ACCESS

Fundamental Challenge

Access to Data Independent of:

- Location
  (File or Database Management System)
- Representation Notation
- Serialization Format
- Transmission Protocol
- Host Operating Systems
- Consumer Applications



OPENLINK SOFTWARE
Making Technology Work For You®

Critical Components

- Identifiers that denote (signify) each entity associated with the following relationship roles:
  - ✓ Entity (Subject)
  - ✓ Attribute (Predicate)
  - ✓ Value (Object)
- Identifiers that denote entity description documents (Descriptors)
- Identifiers that provide entity naming (identification) via implicit or explicit [denotation] → [description document content] resolution using indirection (i.e., combined effect of denotation & connotation to deliver identification or sense)
- Name Resolution Protocols
- Document Content Serialization Formats

Entity Identifiers (Names)

Uniform Resource Identifier (URI)

<http://kingsley.idehen.net/dataspace/person/kidehen#this>
– WebID (i.e., an HTTP URI that denotes an Entity of Type: Agent (Person, Organization, Software, Robot etc)

ODBC Data Source Name (DSN)

DSN=CRM

JDBC Data Source Name (DSN)

DSN=CRM

Entity Description Document Locators

- Uniform Resource (Data) Locator (URL)
  - <http://kingsley.idehen.net/dataspace/person/kidehen> – an HTTP URI that denotes a Document on an HTTP Network
- ODBC Data Source Name
  - `DSN=CRM;HOST=crm.example.org;SVT=Oracle;DATABASE=CRM;TABLE=CUSTOMER` – denotes an ODBC accessible Table in a SQL RDBMS
- JDBC Data Source URL
  - `jdbc:openlink://crm.example.org/SVT=Oracle/DATABASE=CRM/TABLE=CUSTOMER` – denotes a JDBC accessible Table in a SQL RDBMS

ODBC Data Source Name Challenges

- SQL Relational Database Specific.
- Identifiers are x.500 names that are only understood by operating system locked applications.
- Identifiers denote RDBMS specific tables, views, users, and stored procedures.

JDBC Data Source Name Challenges

- SQL Relational Database Specific.
- Identifiers are “jdbc:” scheme URIs that are only understood by JDBC compliant applications constrained by Java Virtual Machine (JVM).
- Identifiers denote RDBMS specific tables, views, users, and stored procedures.

HTTP URI based Data Source Name Virtues

- Database Engine Independent.
- Data Access Protocol Independent.
- Data Representation Format Independent.
- Identifiers are Literals and/or References *(which globalize lookup scope).*
- Identifiers denote anything, i.e., an kind of entity.
- Identifiers are “terms” that resolve to referent description documents, globally.

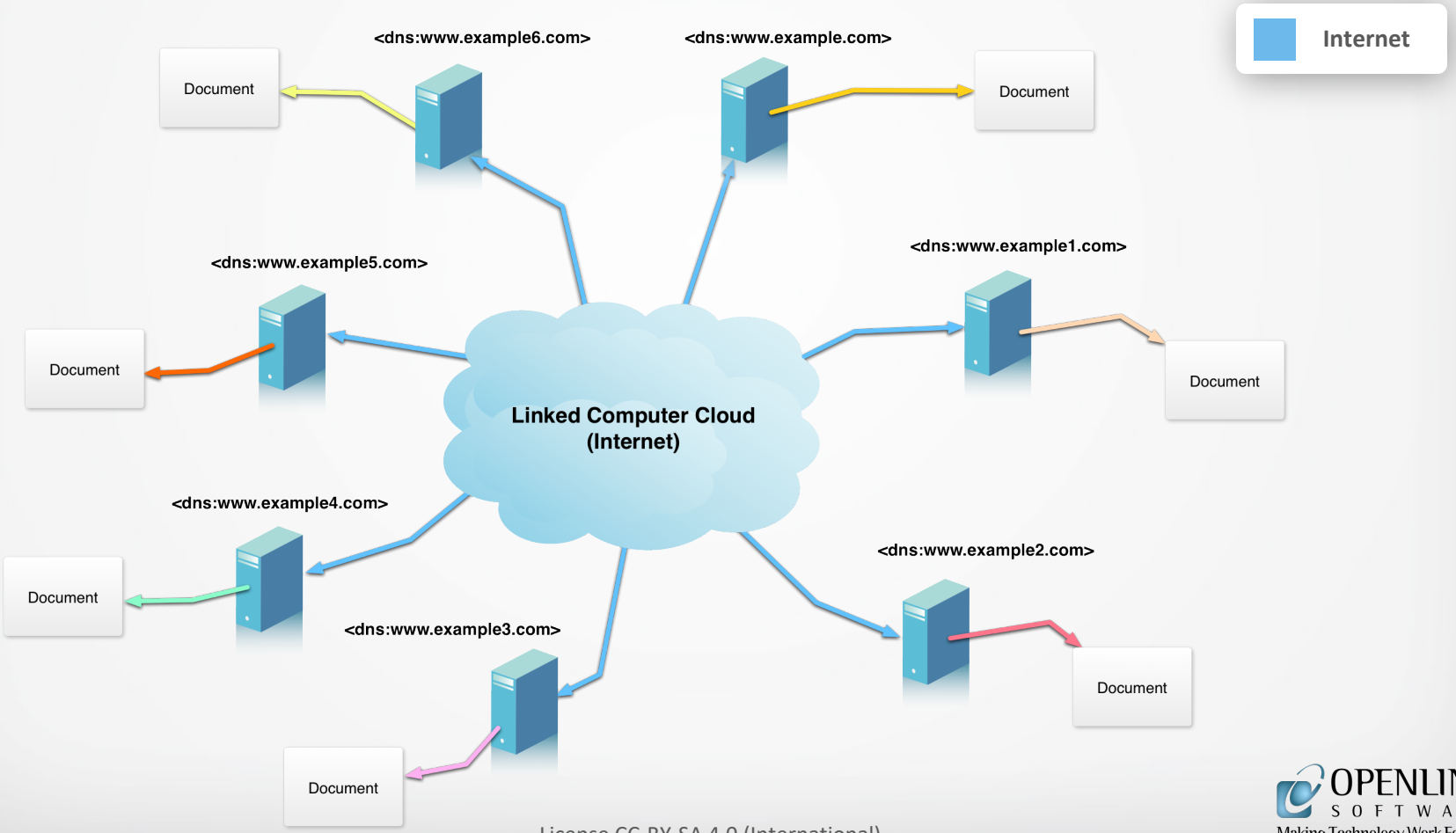
Data Source Name Resolution Protocols

- **Internet based Computer Network** – Domain Name Services (DNS) protocol provides Name Resolution for Computers.
- **World Wide Web Document Network** – HTTP provides Name Resolution for Web Documents via HTTP URLs.
- **World Wide Web Data Network** – HTTP provides Name Resolution for Entities via HTTP URIs .

DNS based Linked Computer Network (Internet)

Linked Computer Network (e.g., Internet)

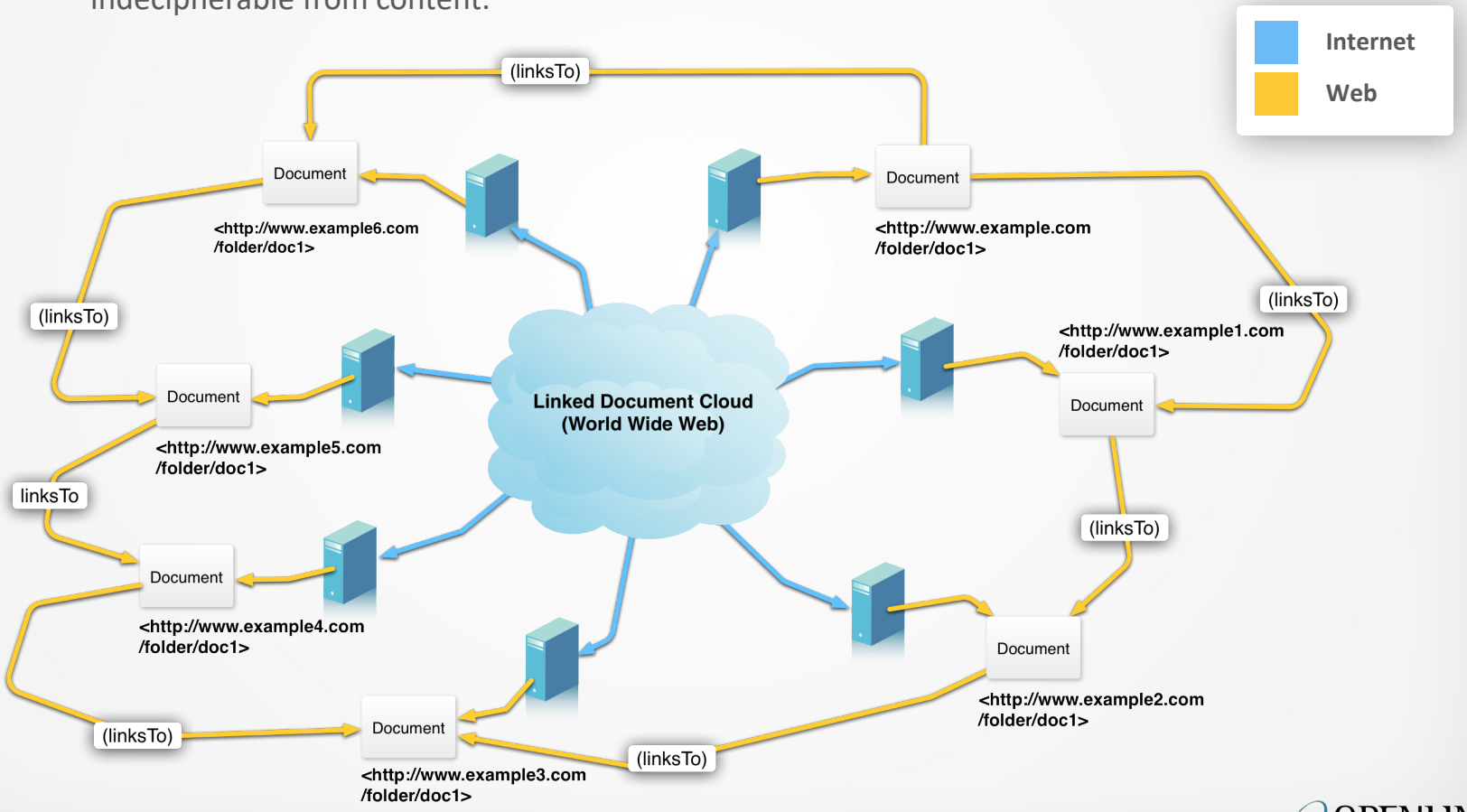
1. Computer (DNS CNAMES) Names are Data Source Name
2. Actual Data Model and Data Access is Local and Machine OS hosted App. specific.



HTTP based Linked Document Network (Web 1.0 & 2.0)

Linked Document Network (e.g., World Wide Web)

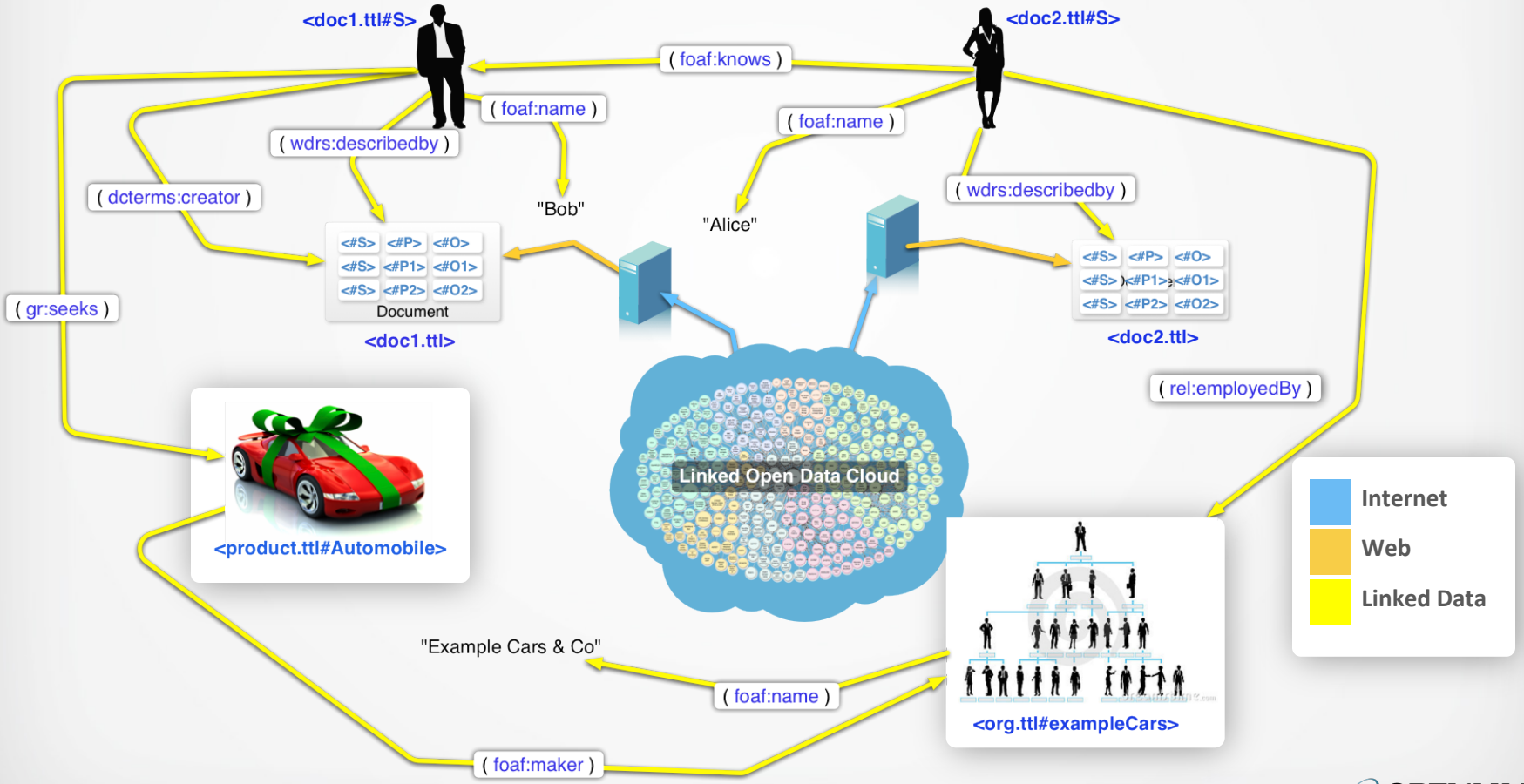
1. Computer (DNS CNAMES) Names become irrelevant.
2. Document Locators / Addresses (HTTP URLs) are Data Source Names (DSNs).
3. One kind of Relation i.e., "LinksTo" is what connects the Documents.
4. To machines: actual Data Model, Entity Relation Semantics, and Representation Notations are indecipherable from content.



HTTP based Linked Data Network (Web 3.0)

Linked Data Network (e.g., Linked Open Data Cloud)

1. Entity Names (HTTP URIs) are Data Source Names (DSNs)
2. Computer (DNS CNAMES) & Document Names (HTTP URLs) become irrelevant
3. Actual Data Model and Representation Notations are loosely coupled.



OPENLINK SOFTWARE
Making Technology Work For You®

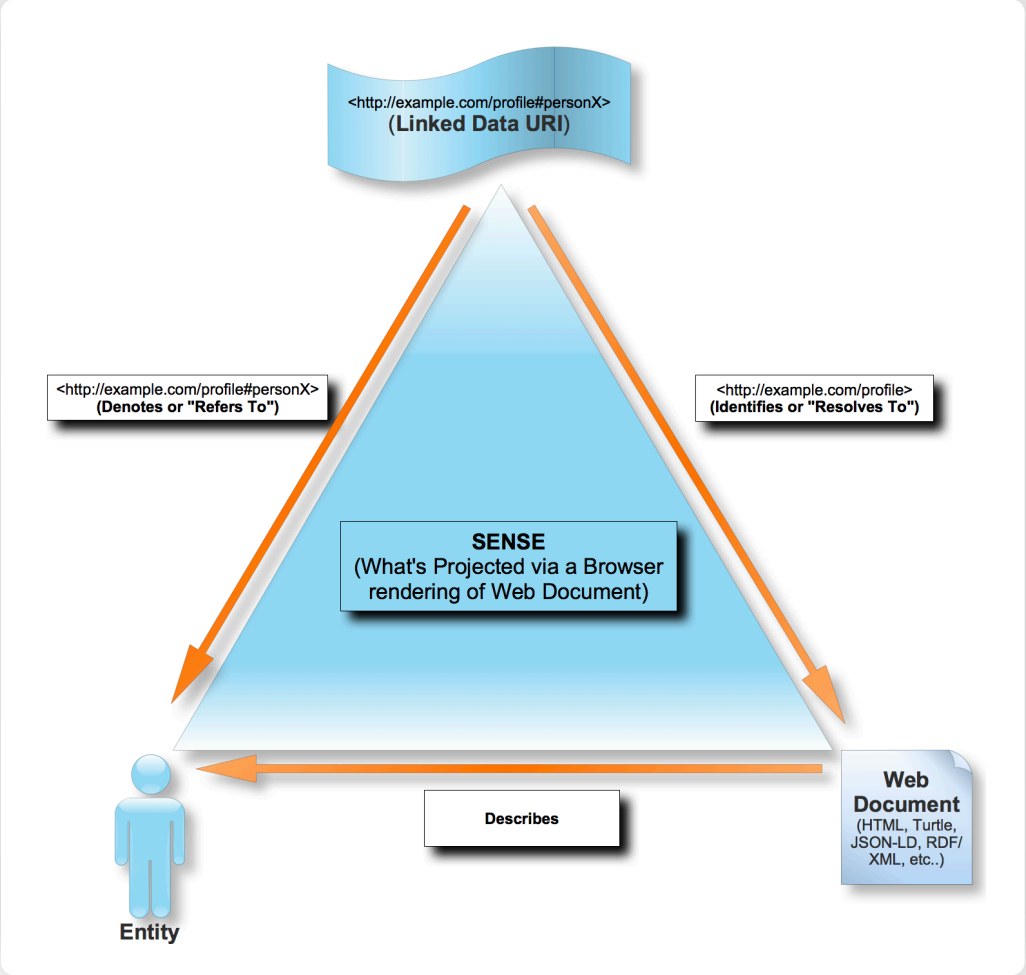
LINKED DATA
(WEBBY STRUCTURED DATA)

Linked Data Fundamentals

- Denote ("refer to" or name) entities unambiguously using URIs – similar to the role of "words" in natural language.
- Use HTTP URIs so that the description of any entity can be looked up using any HTTP user agent – similar to the role of "terms " in natural language.
- Use human and machine readable statements (via open standards e.g., RDF) to create document content that describes entities.
- Refer to other entities using their HTTP URI based names in your entity description documents – i.e., – **expand the Web!**

Understanding HTTP URI Entity Name and Description Doc Address Duality

An HTTP URI is a kind of identifier that denotes (“Refers To”) an entity while also resolving to its description document, over an HTTP Network.



What is Linked Data?

Linked Data is the use of Resolvable URIs
to enhance Structured Data Representation.

Basically:

Representing Entity Relationships using *Statements*
where the relationship role participants
[*Subject*, *Predicate*, and *Object* (optionally)]
are unambiguously "referred to" using *Resolvable URIs*.

What is Linked Open Data?

Linked Open Data is the use of HTTP URIs
to enhance Structured Data Representation.

Basically:

Representing Entity Relationships using *Statements*
where the relationship role participants
[*Subject*, *Predicate*, and *Object* (optionally)]
are unambiguously "referred to" using *HTTP URIs.*

Note: *URIs and HTTP are Open Standards*

Why is Linked Open Data Important?

- It turns HTTP URIs (Hyperlinks) into Data Source Names.
- It moves us from Open Database Connectivity to Open Data Connectivity – that scales from Private Data Spaces to the World Wide Web.
- It delivers a powerful mechanism for virtualization of disparate and heterogeneous data sources (big or small) i.e., Data De-Silo-Fication.
- It is inherently Platform Agnostic.
- It delivers a Linked Open Data Cloud that scales to the World Wide Web.

What is RDF based Linked Data?

RDF-based Linked Data is the use of IRIs and Entity Relationship Type (aka. Relations) Semantics to enhance Structured Data Representation.

Basically:

Representing Entity Relationships using *Statements* where the relationship role participants [*Subject*, *Predicate*, and *Object* (optionally)] are unambiguously "referred to" using *IRIs*.

Note: *RDF and IRIs are Open Standards*

What is RDF based Linked Open Data?

RDF-based Linked Open Data is the use of *HTTP URIs* & Entity *Relationship Type (Relations) Semantics*
to enhance Structured Data Representation.

Basically:

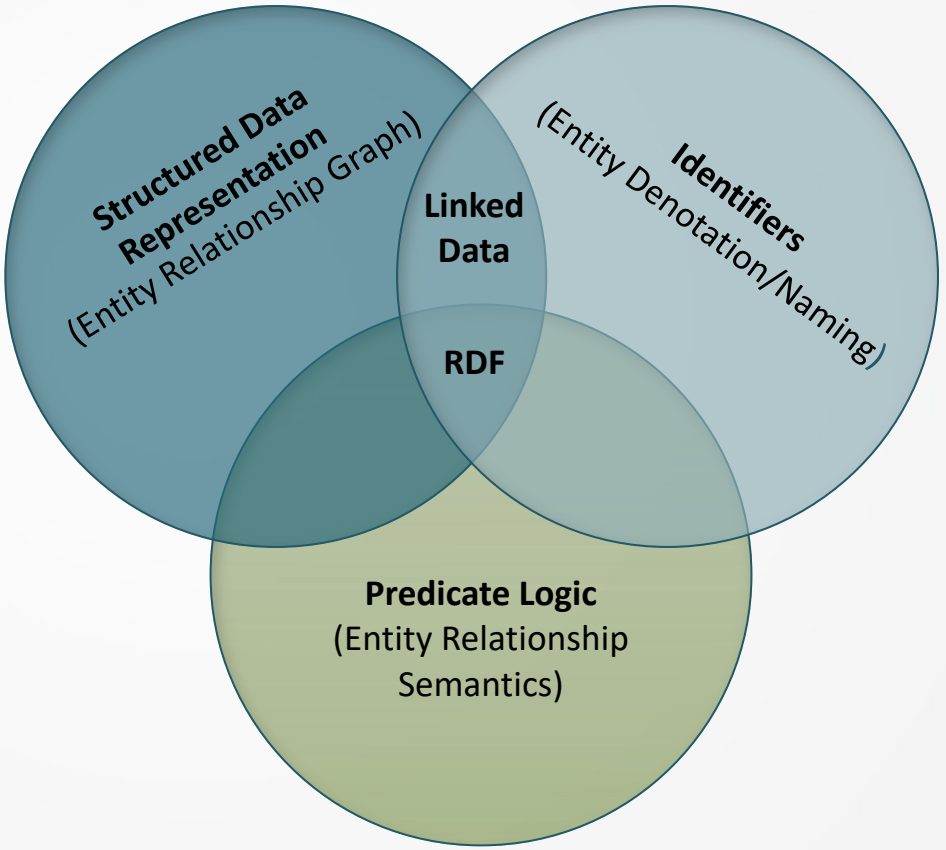
Representing Entity Relationships and *Relation Semantics* using *Statements*
where the relationship role participants
[*Subject*, *Predicate*, and *Object* (optionally)]
are unambiguously "referred to" using *HTTP URIs.*

Note: *RDF, HTTP and URIs are Open Standards*

What is RDF based Linked Data?

RDF-based Linked Data is Web-Like Structured Data enhanced with RDF's **explicit** machine-and human-comprehensible Entity Relationship Semantics.

Identifiers, Structured Data Representation, and Logic

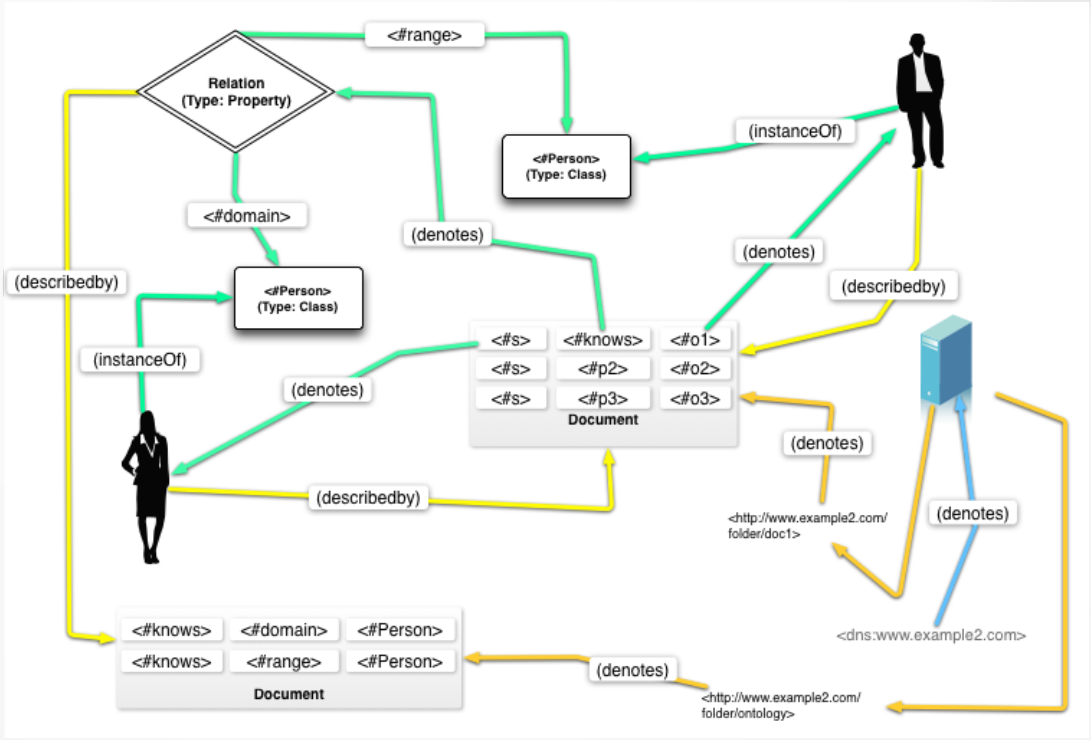


RDF based Linked Open Data (Semantic Web)

Semantically Enhanced Linked Data Network (e.g., Semantic Web of Big Linked Open Data)

- Internet
- Web
- Linked Data
- Relation Semantics

1. Entity Names (HTTP URIs) are Data Source Names (DSNs)
2. Computer (DNS CNAMES) & Document Names (HTTP URLs) become irrelevant
3. Actual Data Model and Representation Notations are loosely coupled
4. **RDF** & **RDF Schema** Relation Semantics are accessible and comprehendible to humans and machines.



OPENLINK SOFTWARE
Making Technology Work For You®

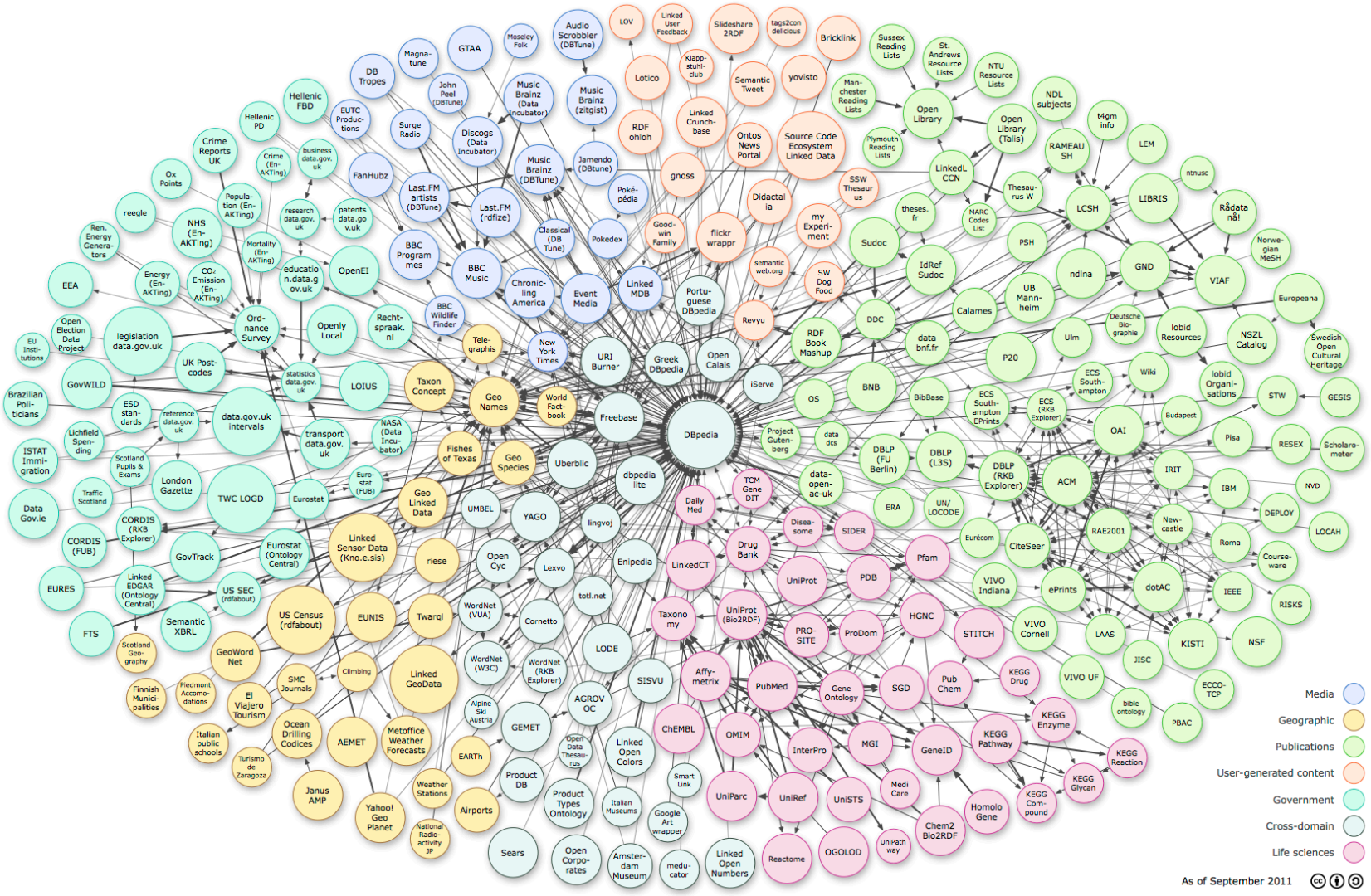
Local Linked Data (Inaccessible)

Entity (Subject)	Attribute (Predicate)	Value (Object)
urn:data:object:id:France	urn:data:object:id:Type	urn:data:object:id:PopulatedPlace
urn:data:object:id:France	urn:data:object:id:hasLabel	“France”
urn:data:object:id:France	urn:data:object:id:hasCapital	urn:data:object:id:Paris
urn:data:object:id:Paris	urn:data:object:id:Type	urn:data:object:id:PopulatedPlace
urn:data:object:id:Paris	urn:data:object:id:hasLabel	“Paris”
urn:data:object:id:PopulatedPlace	urn:data:object:id:Type	urn:data:object:id:Place

Linked Data (Accessible Webby Data)

Entity (Subject)	Attribute (Predicate)	Value (Object)
http://dbpedia.org/resource/France	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/PopulatedPlace
http://dbpedia.org/resource/France	http://www.w3.org/2000/01/rdf-schema#label	“France”
http://dbpedia.org/resource/France	http://dbpedia.org/ontology/capital	http://dbpedia.org/resource/Paris
http://dbpedia.org/resource/Paris	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/PopulatedPlace
http://dbpedia.org/resource/Paris	http://www.w3.org/2000/01/rdf-schema#label	“Paris”
http://dbpedia.org/ontology/PopulatedPlace	http://www.w3.org/2000/01/rdf-schema#subClassOf	http://dbpedia.org/ontology/Place

Massive Linked Open Data Cloud



NATURAL LANGUAGE & DATA

"Natural Languages are the most sophisticated systems of communication ever developed." – John F. Sowa

"Once you have a truly massive amount of information integrated as knowledge, then the human-software system will be superhuman, in the same sense that mankind with writing is superhuman compared to mankind before writing." – Douglas Lenat

Natural Language & Data

- A Word or Phrase is an identifier that names an Entity (thing) via implicit [denotation]→[referent description document content] resolution
- A Term is a Word or Phrase that names an Entity via explicit, [denotation]→[referent description document content] resolution, using indirection.
- A Sentence is a syntax rules constrained arrangement of Words and Phrases that represent types of Entity Relationships.
- A Statement is a kind of Sentence constructed from Terms.

Data (Recap)

- A IRI is an Internationalized Identifier that has the entity naming characteristics of a Word or Phrase.
- An HTTP URI is a kind of IRI that has the entity naming characteristics of a Term i.e., denotation (signification) and connotation (description) reference duality.
- RDF enables digital sentence construction where IRIs are used to name Entities participating in the Subject, Predicate, and Object relationship roles.
- RDF based Linked Data enables digital statement construction where HTTP URIs are used to denote Entities participating in the Subject, Predicate, and Object relationship roles.

Natural Language & Data Connection

- An RDF triple represents a "Datum" – a Sentence compromised of Words or Phrases.
- An RDF based Linked Open Data Triple represents a "Webby Datum" – a Statement comprised of Terms.
- RDF triple collections represent Data – Sentences.
- RDF based Linked Open Data triple collections represent "Webby Data" – Statements.

Live Additional Information Links

An Glossary of terms, in Linked Data form:

- Data
- Big Data
- Open Data
- Public Open Data
- Linked Data
- Linked Open Data
- Semantic Web
- Resource Description Framework (RDF)

References

- The Role of Logic and Ontology in Language and Reasoning --- John F. Sowa
- Blogic – Pat Hayes
- Unified View of Data – Peter Chen
- Levels of Abstraction: Net, Web, Graph – Tim Berners-Lee
- What is Data? What is a Datum – Ontolog Forum Thread
- Data & Relations – Ontolog Forum Thread.

Additional Information

Web Sites

OpenLink Software

YouID – Digital Identity Card (Certificate) Generator

OpenLink Data Spaces – Semantically enhanced Personal & Enterprise Data Spaces & Collaboration Platform

OpenLink Virtuoso - Hybrid Data Management, Integration, Application, and Identity Server

Universal Data Access Drivers - High-Performance ODBC, JDBC, ADO.NET, and OLE-DB Drivers

Social Media Data spaces

http://kidehen.blogspot.com *(weblog)*

http://www.openlinksw.com/blog/~kidehen/ *(weblog)*

https://plus.google.com/112399767740508618350/posts *(Google+)*

https://twitter.com/#!/kidehen *(Twitter)*

Hashtag: #LinkedData *(Anywhere).*