

UNIVERSITY OF  
Southampton

# Linked and Open Data

COMP3220 Web Infrastructure

Dr Nicholas Gibbins – [nmg@ecs.soton.ac.uk](mailto:nmg@ecs.soton.ac.uk)



A goal of the Web was that, if the interaction between person and hypertext could be so intuitive that the machine-readable information space gave an accurate representation of the state of people's thoughts, interactions, and work patterns, then machine analysis could become a very powerful management tool, seeing patterns in our work and facilitating our working together



I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines.

# What is the Semantic Web?

“The goal of the Semantic Web initiative is as broad as that of the Web: to create a universal medium for the exchange of data. It is envisaged to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific and cultural data. Facilities to put machine-understandable data on the Web are quickly becoming a high priority for many organizations, individuals and communities.

The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people.”



THE  
SEMANTIC  
WEB

# The annotated Web

Add structure to unstructured data

- Annotate existing web pages
- Classify web pages
- Use natural language techniques to extract information from web pages

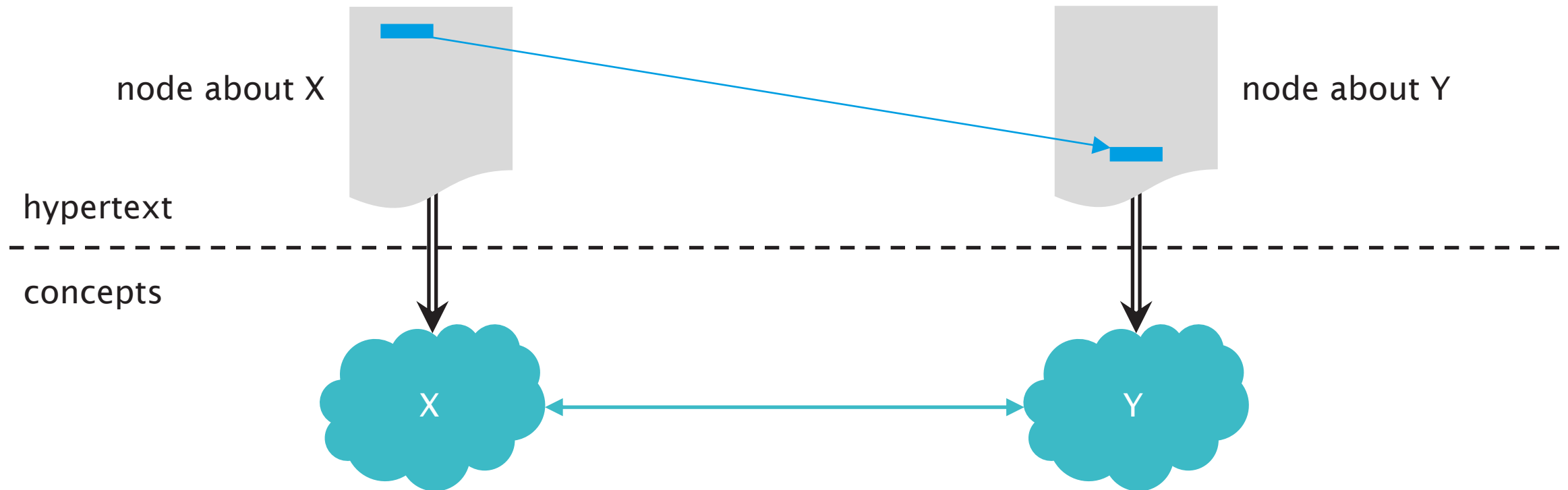
Annotations enable enhanced browsing and searching

Examples: COHSE, Magpie


The image shows a composite screenshot illustrating web annotation. At the top, a Microsoft Internet Explorer window displays a NASA GISS article titled "A Stratospheric 'Clock' to Measure Upper Atmosphere Circulation". The text on the page is annotated with green boxes around terms like "atmosphere", "stratosphere", "troposphere", "aerosol", "precipitation", and "midlatitude". To the right of the text is a heatmap showing atmospheric circulation patterns, with two black circles highlighting specific regions. Below the main text, a sidebar titled "COHSE DLS" is visible, containing a "Settings" button and a "Link Status" section. The "Link Status" section shows "Added 27 (from 101) generic links" and "Added 3 (from 3) annotation links". At the bottom of the screenshot, a Mozilla browser window displays a tutorial page titled "Writing Filters for Random Access Files". This page also features annotations, such as a blue box around the text "The example CheckedIODemo from How to Write Your Own Filter Streams implements two filter streams that compute a checksum as data is read from or written to the stream." and a small dialog box titled "Checksum" that explains the concept of a checksum.



# Hypertext and linked data





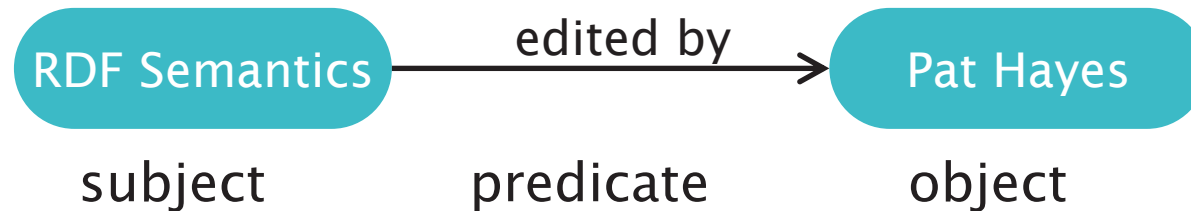
A close-up photograph of a middle-aged man with short, graying hair, wearing a blue button-down shirt. He is looking slightly to his right and speaking into a silver microphone. The background is dark and out of focus.

Is this rocket science? Well, not really. The Semantic Web, like the World Wide Web, is just taking well established ideas, and making them work interoperably over the Internet. This is done with standards, which is what the World Wide Web Consortium is all about. We are not inventing relational models for data, or query systems or rule-based systems. We are just webizing them.

# Resource Description Framework

Underlying model of triples used to describe the relations between entities

- Subject-Predicate-Object (compare Entity-Attribute-Value)
- Predicates are analogous to link types



# Example

Take a citation:

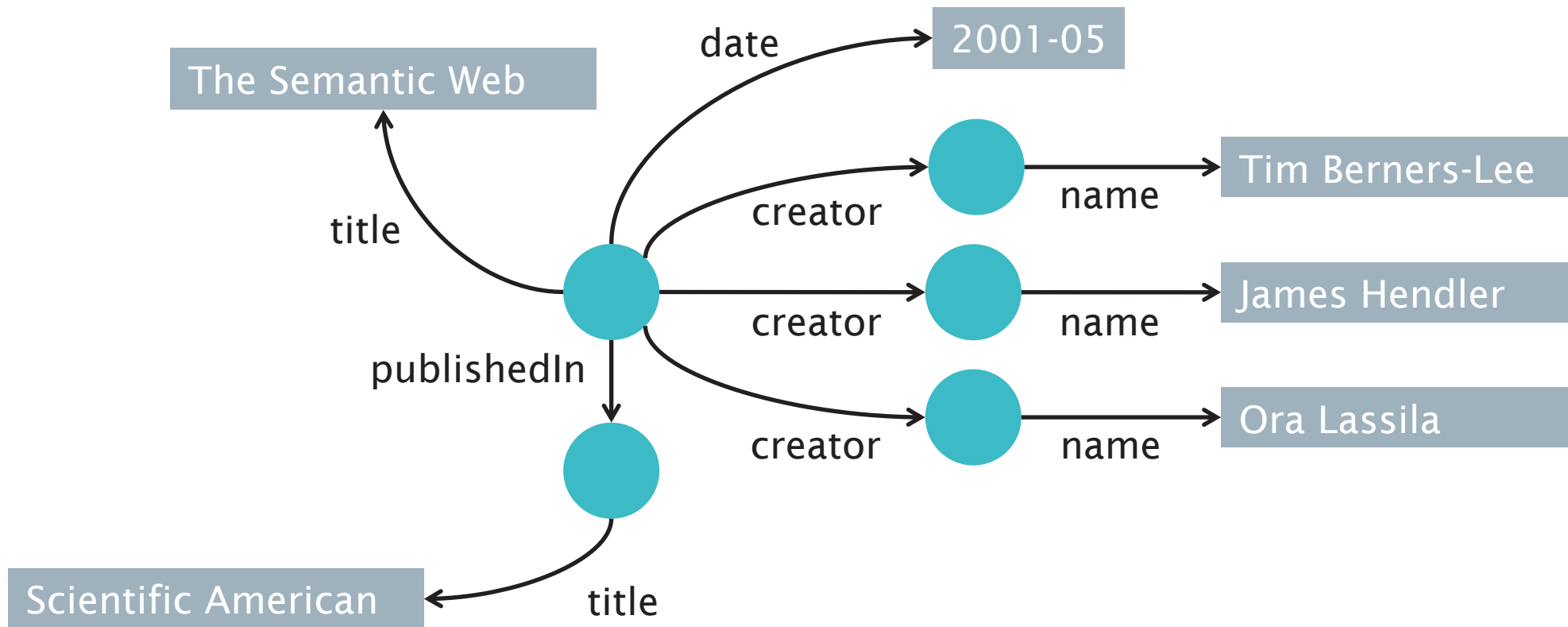
- Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web. Scientific American, May 2001

We can identify a number of distinct statements in this citation:

- There is an article titled “The Semantic Web”
- One of its authors is a person named “Tim Berners-Lee” (etc)
- It appeared in a publication titled “Scientific American”
- It was published in May 2001

# Example

We can represent these statements as a graph:



# Example

There are two types of node in this graph:

- **Literals**, which have a value but no identity  
(a string, a number, a date)

Scientific American

- **Resources**, which represent objects with identity  
(a web page, a person, a journal)

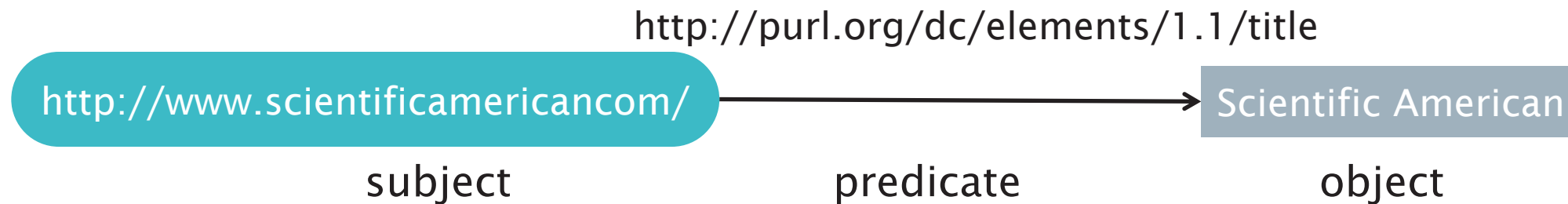


# Example

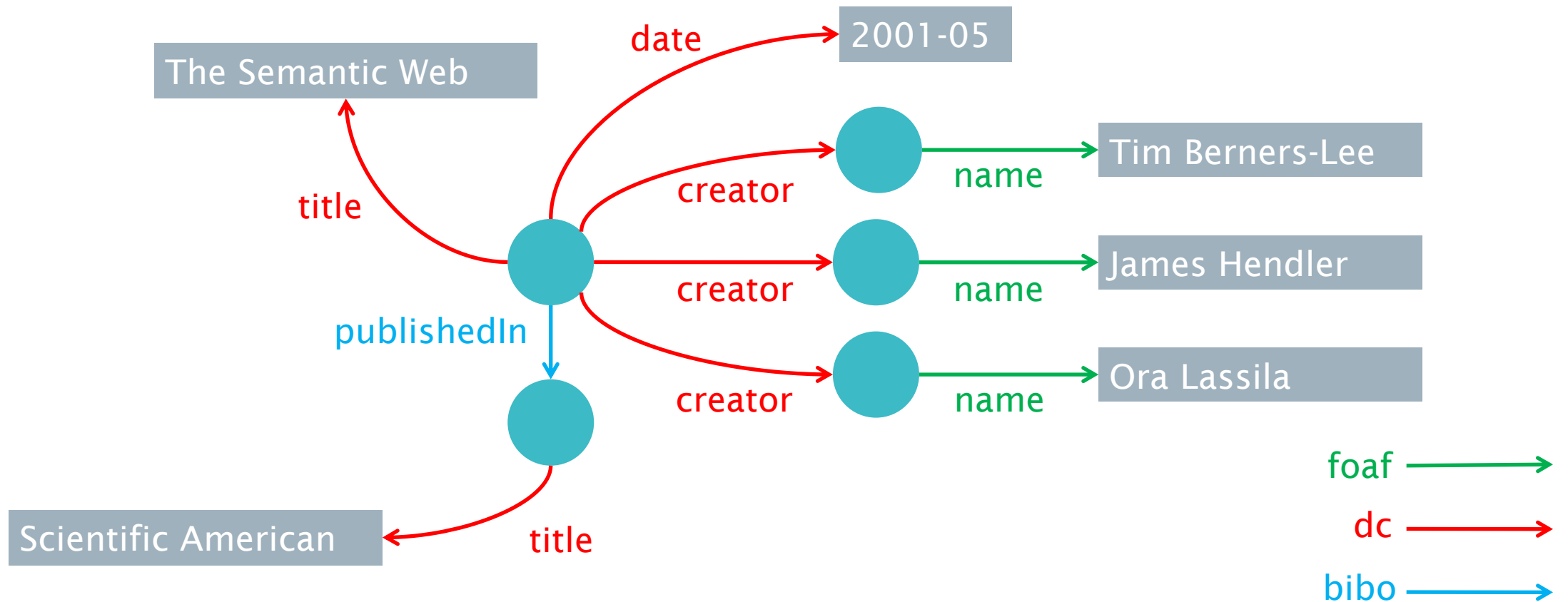
Resources are identified by URIs

Properties are resources that are used as predicates

- Collection of properties constitutes a vocabulary (or ontology)



# Mixing vocabularies



# Linked data principles

Set of publishing practices for Semantic Web data:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information
4. Include links to other URIs so that they can discover more things



# 1. Use URIs as names for things

Use URIs as unique identifiers to denote things:

- Hegel, Georg Wilhelm Friedrich
  - [http://dbpedia.org/resource/Georg\\_Wilhelm\\_Friedrich\\_Hegel](http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel)
  - <http://viaf.org/viaf/89774942>
  - ...
- Hegel, Georg Wilhelm Friedrich: *Gesammelte Werke / Vorlesungen über die Logik*
  - <urn:isbn:978-3-7873-1964-0>



## 2. Use HTTP URIs

Enables “lookup” of URIs via HTTP

- (i.e. fetch a resource representation that provides a description of a thing)
- Uses established infrastructure – piggy-backs on Domain Name System

Connects logical level (thing) with physical level (source)

Important distinction between names for things and locations of sources

- *Non-information resources versus information resources*
- See also TAG issue httpRange-14 (beware can of worms!)

## 3. Provide useful information

When somebody looks up a URI, return data using the standards (RDF\*, SPARQL)

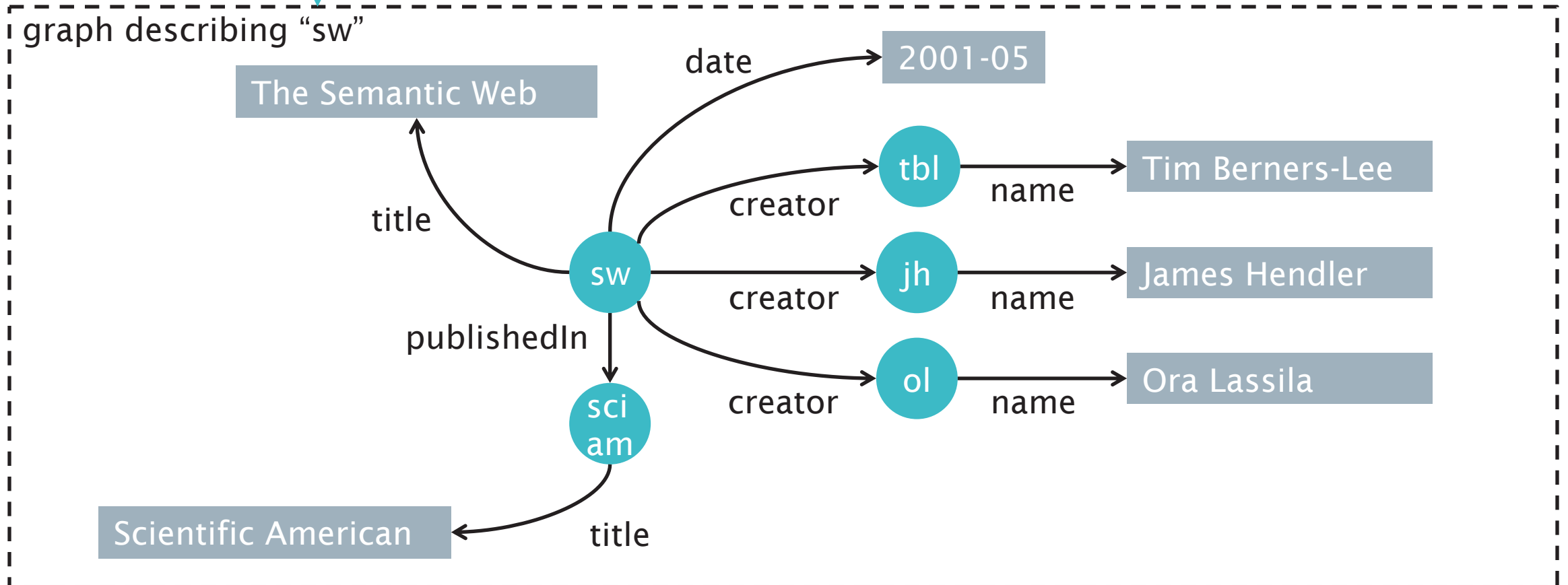
Representation = encoding of information about resource state

Representations as descriptions of resources

sw

(short for <https://www.scientificamerican.com/article/the-semantic-web/> )

yields on dereference



## 4. Link to other URIs

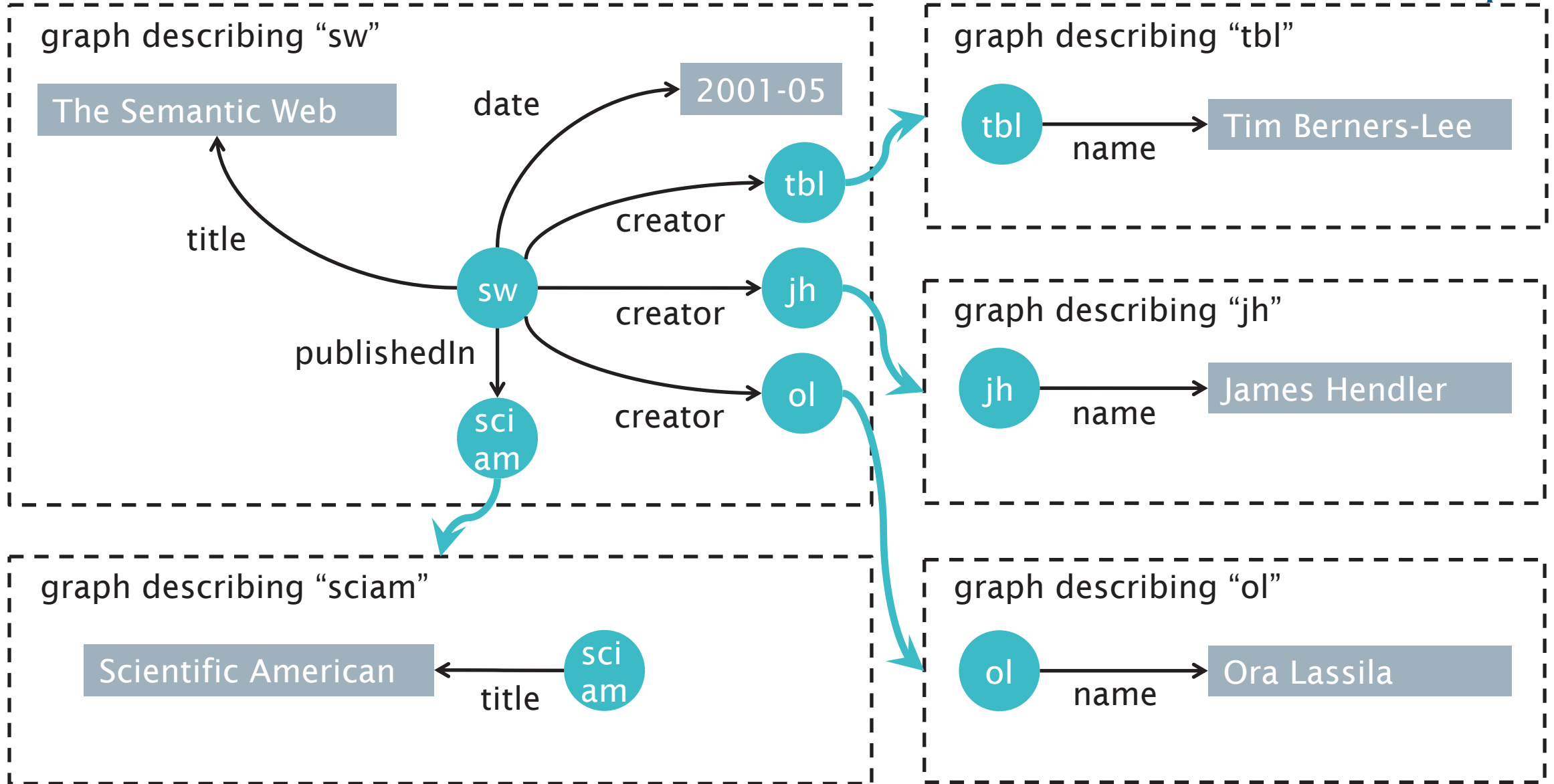
Enable people (and machines) to navigate between resources and datasets

Using external vocabularies enables linking

- What does this property mean? Fetch the property for a definition
- Vocabularies can be interlinked (vocabularies build on other vocabularies)

Special vocabulary to denote equivalence of identifiers (useful for data merging)

- `owl:sameAs`



# Linked data principles

Set of publishing practices for Semantic Web data:

1. Use URIs as names for things
  2. Use HTTP URIs so that people can look up those names
  3. When someone looks up a URI, provide useful information
  4. **Include links to other URIs so that they can discover more things**
- Putting the hypertext back into the Semantic Web
  - Simplifies integration between datasets while maintaining loose coupling

## 5 Stars of Linked Data (2010)

- ★ Available on the Web (in whatever format) under an open licence
- ★★ As above, but as machine-readable structured data (e.g. Excel instead of an image of a table)
- ★★★ As above, but in a non-proprietary format (e.g. CSV instead of Excel)
- ★★★★ As above, but using W3C standards (RDF, SPARQL) to identify things, so that others can point at your data
- ★★★★★ As above, but linked to other people's data to provide context



# From Linked Data to Open Data

Often conflated, but fundamentally different (if complementary)

- Linked Data is about **how data is published online**
- Open Data is about **what users of the data are allowed to do with it**

Open Data is:

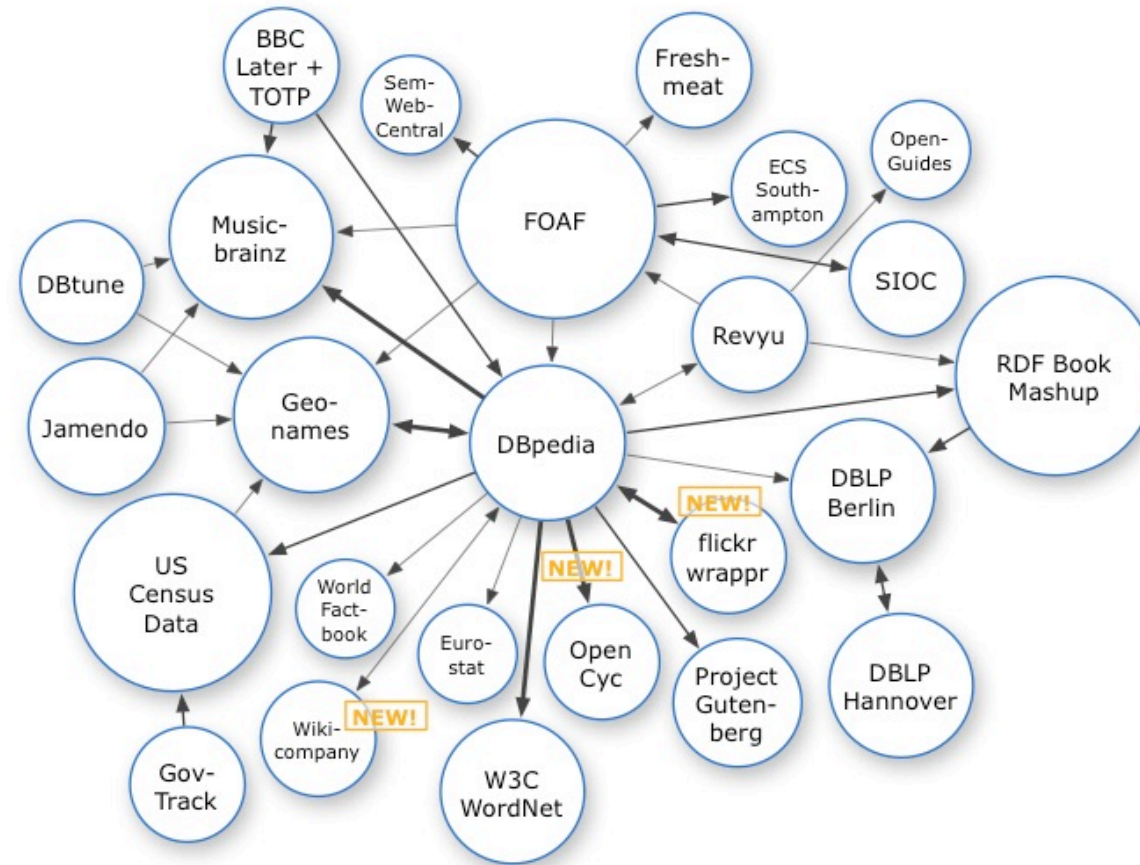
- Freely available to everyone
- To use
- To modify
- To share
- For any purpose

# Open Data licences

- Open licences grant rights to users
  - Public domain licences have no restrictions on use or reuse
  - Attribution licences require reusers to attribute the source of the content
  - Attribution and share-alike licences require reusers to attribute the source of the content and share any derived content under the same licence
- Other open licences are available (e.g. UK Open Government Licence  $\approx$  attribution)

Level of Licence	Creative Commons	Open Data Commons
Public Domain	CC0	PDDL
Attribution	CC-by	ODC-by
Attribution and share-alike	CC-by-sa	ODbL

# Linked Open Data on the Web: 2007

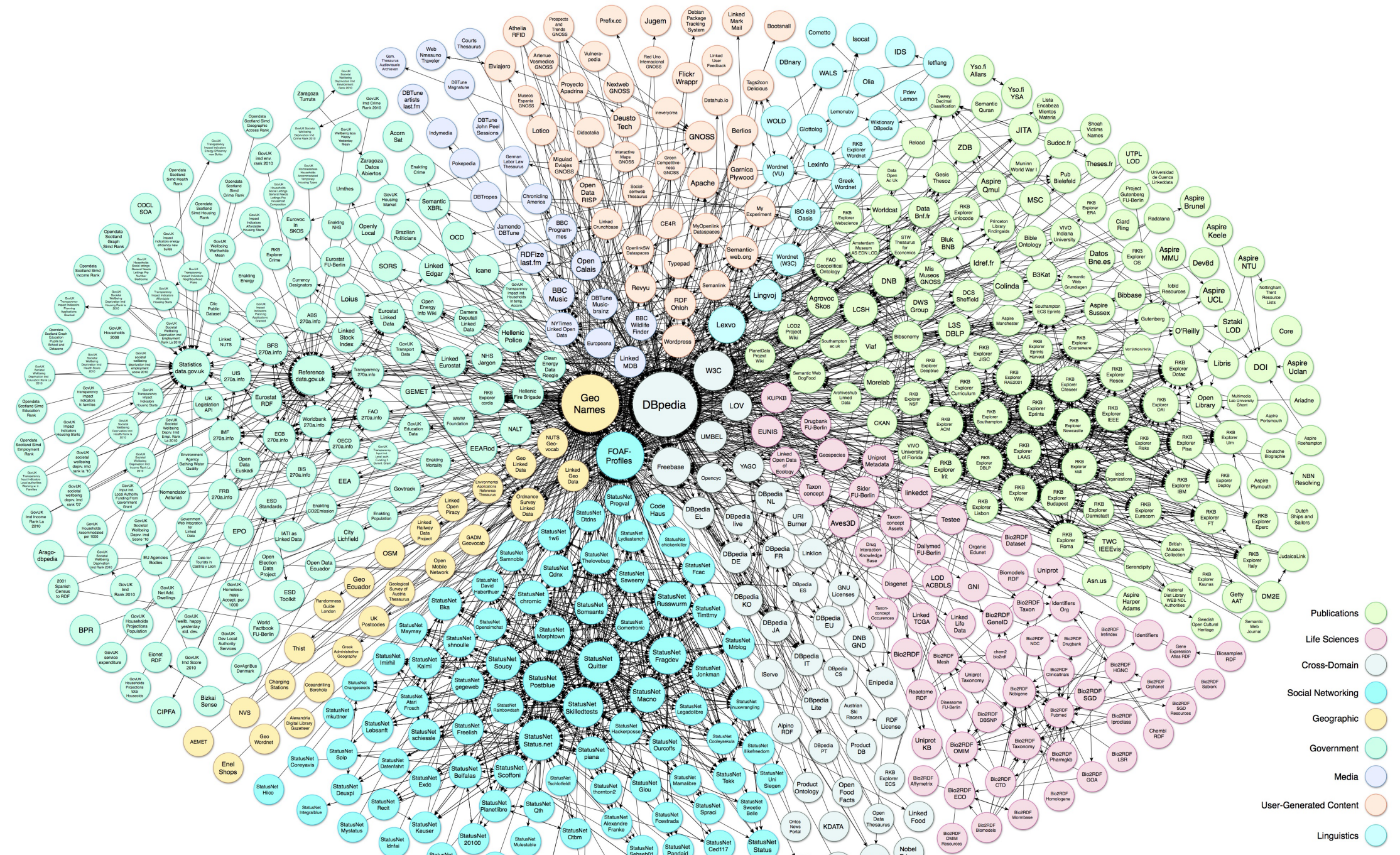








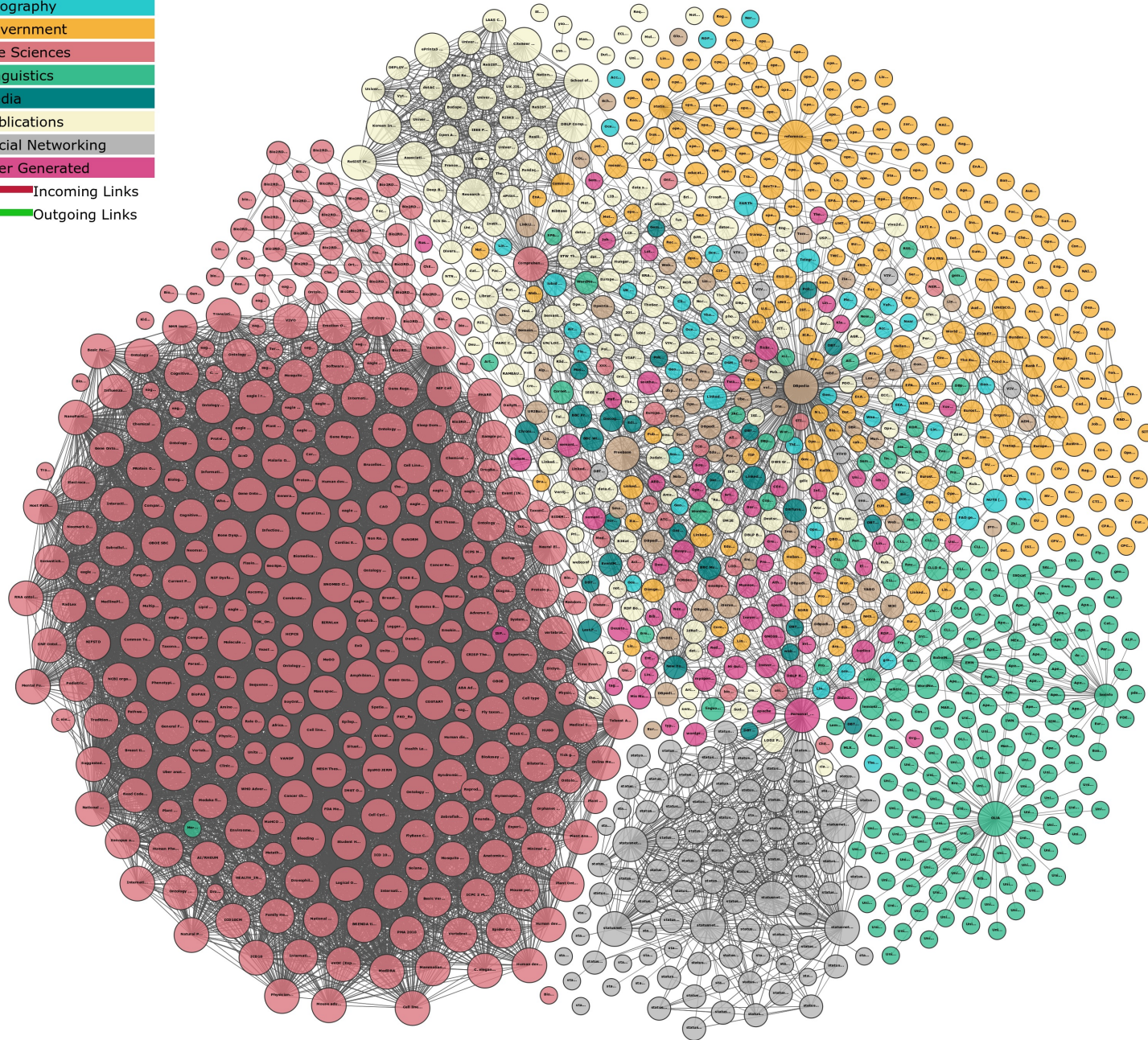
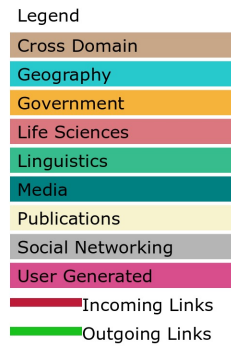




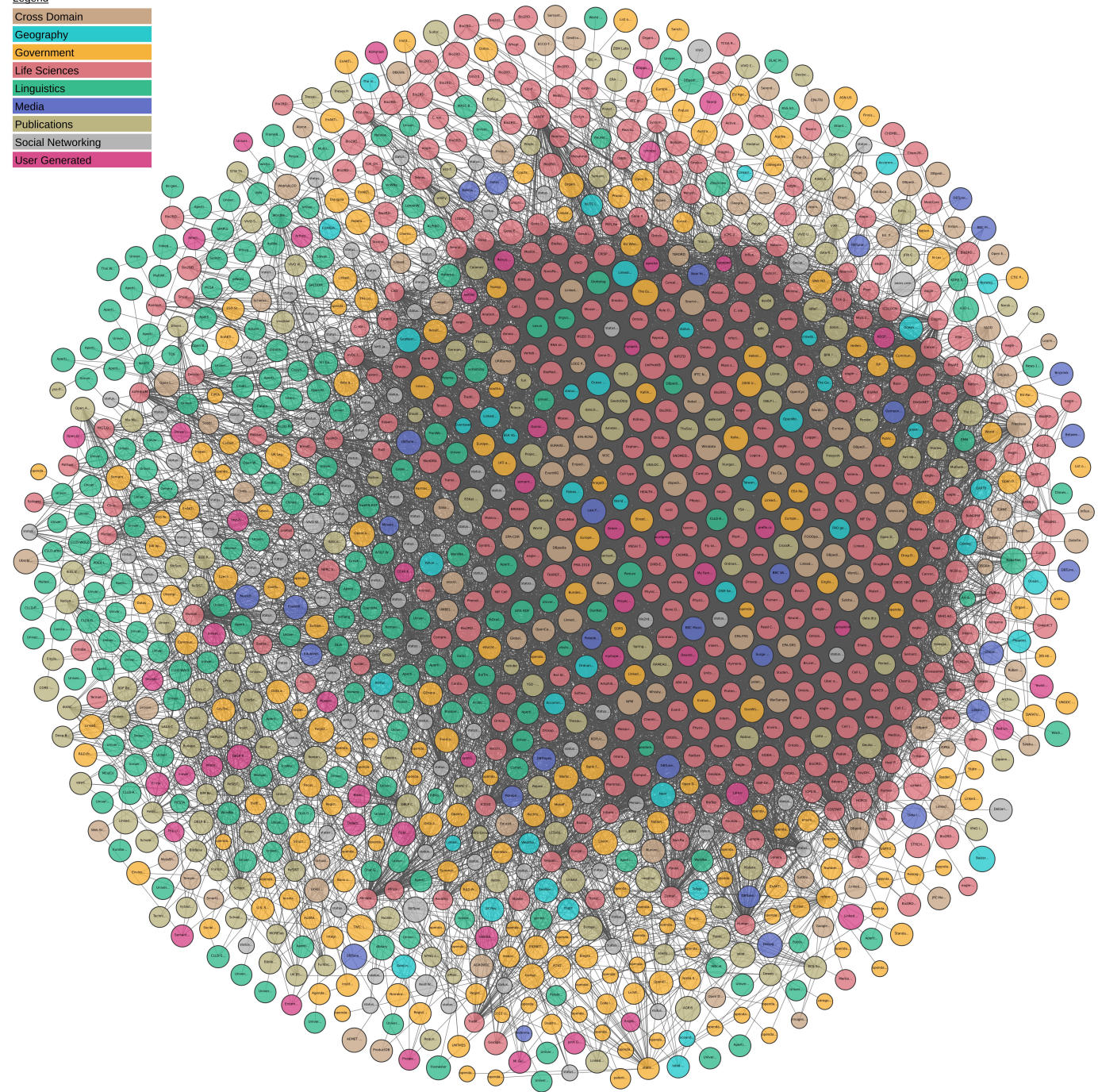
- Publications
- Life Sciences
- Cross-Domain
- Social Networking
- Geographic
- Government
- Media
- User-Generated Content
- Linguistics







- Legend
- Cross Domain
  - Geography
  - Government
  - Life Sciences
  - Linguistics
  - Media
  - Publications
  - Social Networking
  - User Generated



## Further Reading

Berners-Lee, T. et al (2001) *The Semantic Web*. Scientific American. 284(5), pp. 29-37.

Shadbolt, N. et al (2006) *The Semantic Web Revisited*. IEEE Intelligent Systems 21(3) pp. 96-101.

Bernstein, A. et al (2016) *A New Look at the Semantic Web*. Communications of the ACM. 59(9), pp. 35-37.

Next Lecture: Intellectual Property