USC Viterbi
School of Engineering

Linked Data and the Semantic Web

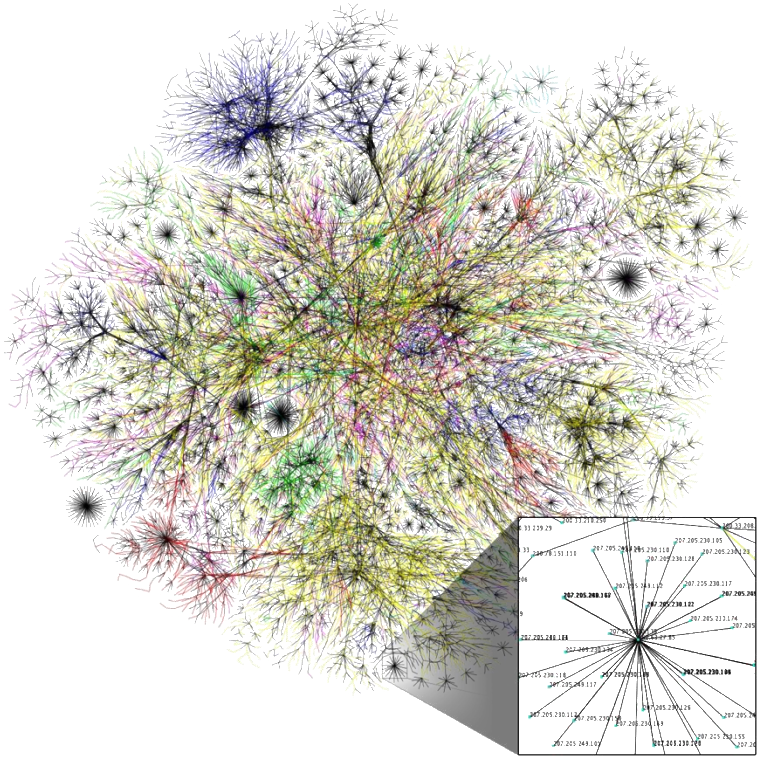
Basel Shbita

DSCI 558: Building Knowledge Graphs

Fall 2020, University of Southern California

The Internet

- 1962, Rand Corporation, communications systems for military
- 1965, Packet switching at NPL
- 1969, ARPANET
- 1971, First Email
- 1981, CSNET & BITNET are born
- 1983, DNS is born
- 1983, TCP/IP is born
- 1985, FTP is standardized

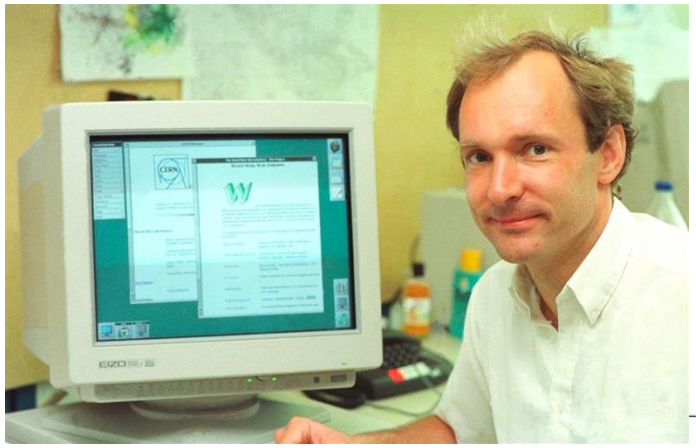


slide by Basel Shbita, picture by the Opte Project https://en.wikipedia.org/wiki/History_of_the_Internet#/media/File:Internet_map_1024_-_transparent,_inverted.png

The World Wide Web

- 1980s, Tim Berners-Lee @ CERN
- 1989, HTTP
- 1990, WWW Proposal
- 1991, HTML
- 1993, Mosaic (NCSA), 1st Web Browser

- *Since*, Web of Documents
- *Now*, Web of Data



Tim Berners-Lee, CERN/DD

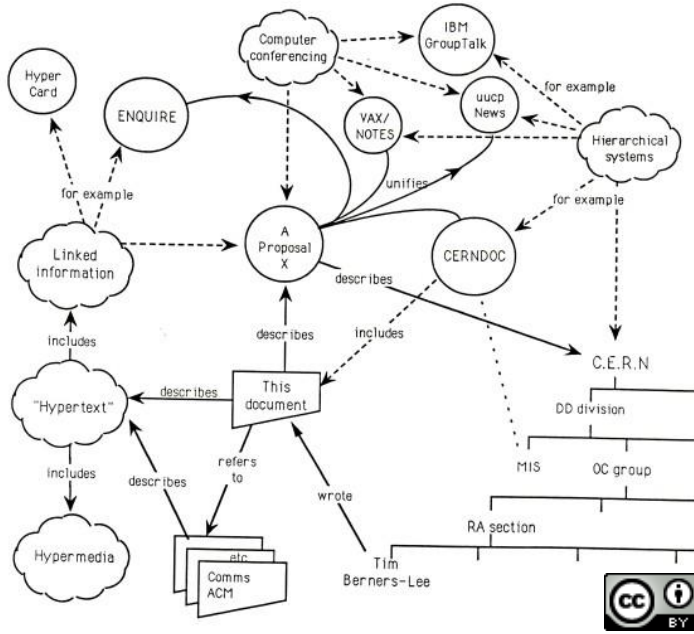
March 1989

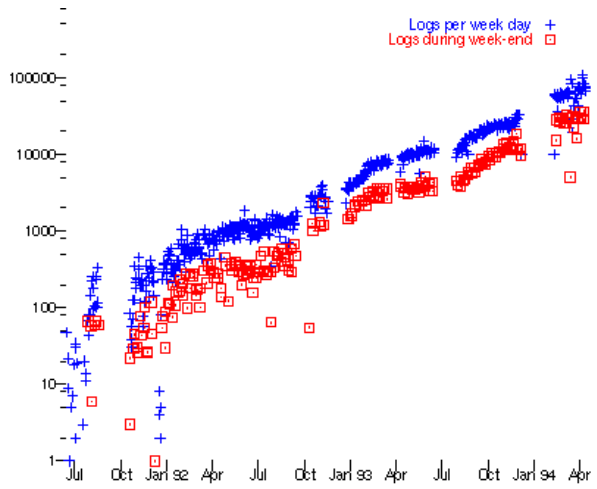
Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control



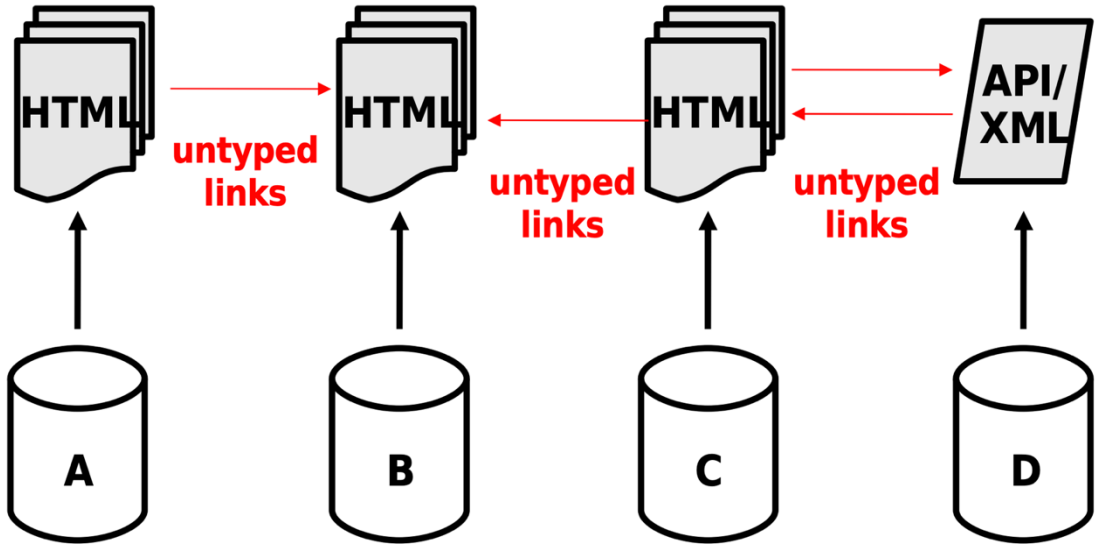


https://www.w3.org/History/1989/proposal.html

slide by Basel Shbita, content from https://www.w3.org/History.html

The Web of Documents

- Analogy
  - a global filesystem
- Primary objects
  - documents
- Links between
  - documents (or sub-parts of)
- Degree of structure in objects
  - fairly low
- Semantics of content and links
  - implicit
- Designed for
  - human consumption

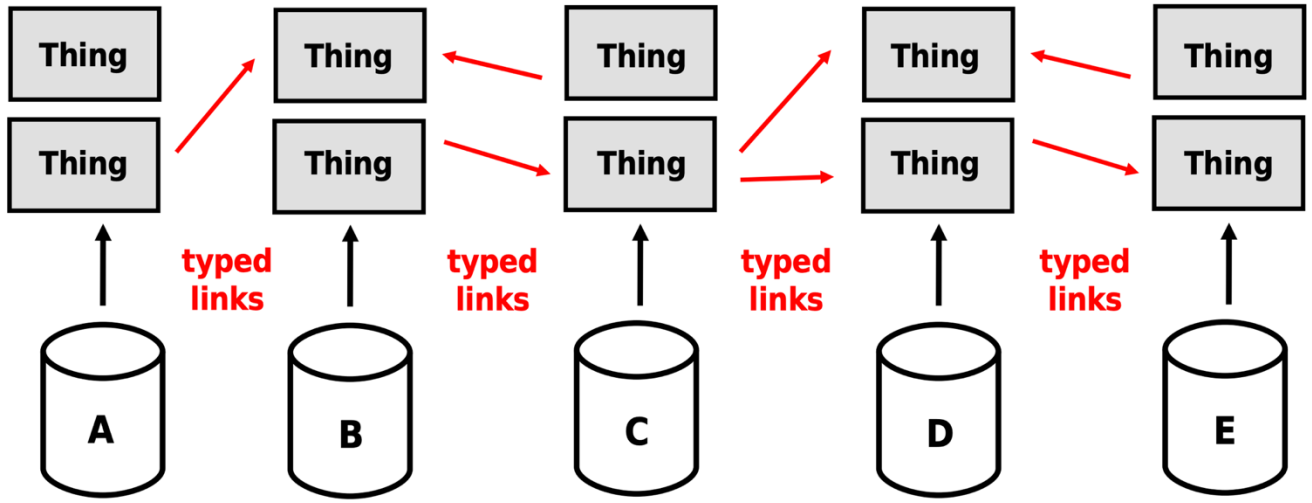


Disconnected Data ☹

slide by Christian Bizer, Tom Heath, Tim Berners-Lee: Linking Open Data (2/19/2008)

The Web of Data

- Analogy
  - a global database
- Primary objects
  - things (or descriptions of things)
- Links between
  - things (including documents)



- Degree of structure in (descriptions of) things
  - high
- Semantics of content and links
  - explicit
- Designed for
  - machines first, humans later

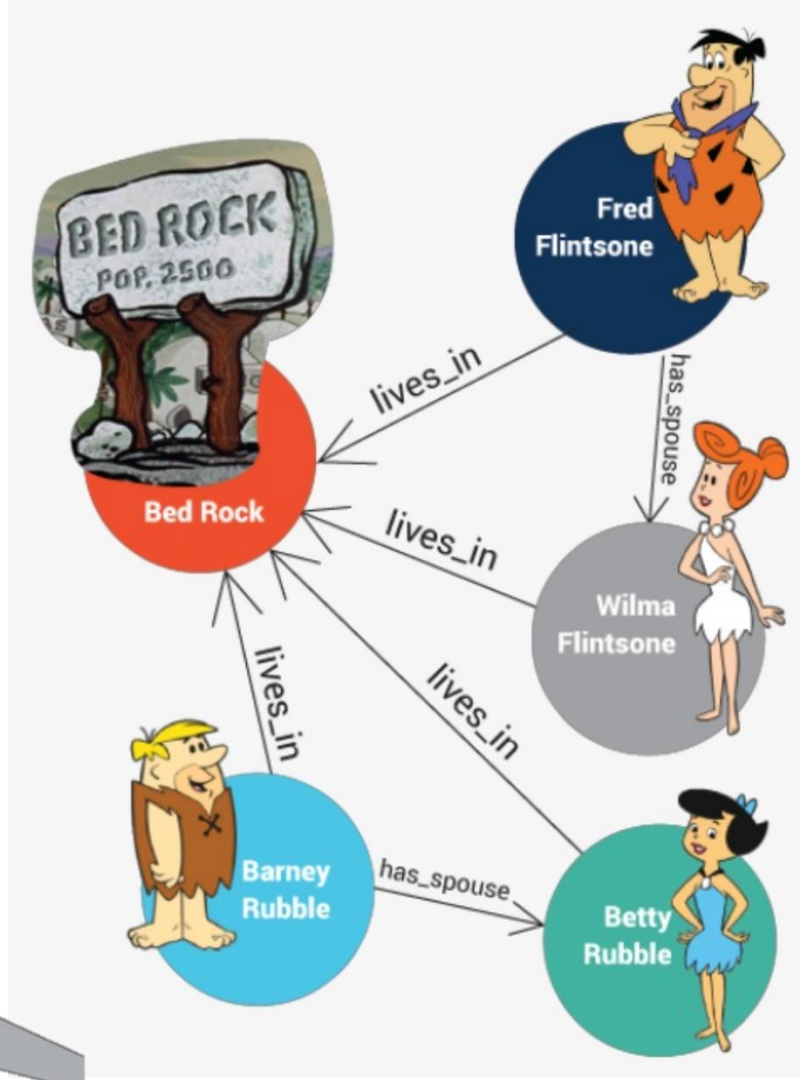
Linked Data

- A method of publishing structured data so that it can be interlinked and become more useful
- Web technologies:
  - HTTP
  - URIs
  - RDF

  to share information → processed automatically by computers

Semantic Web

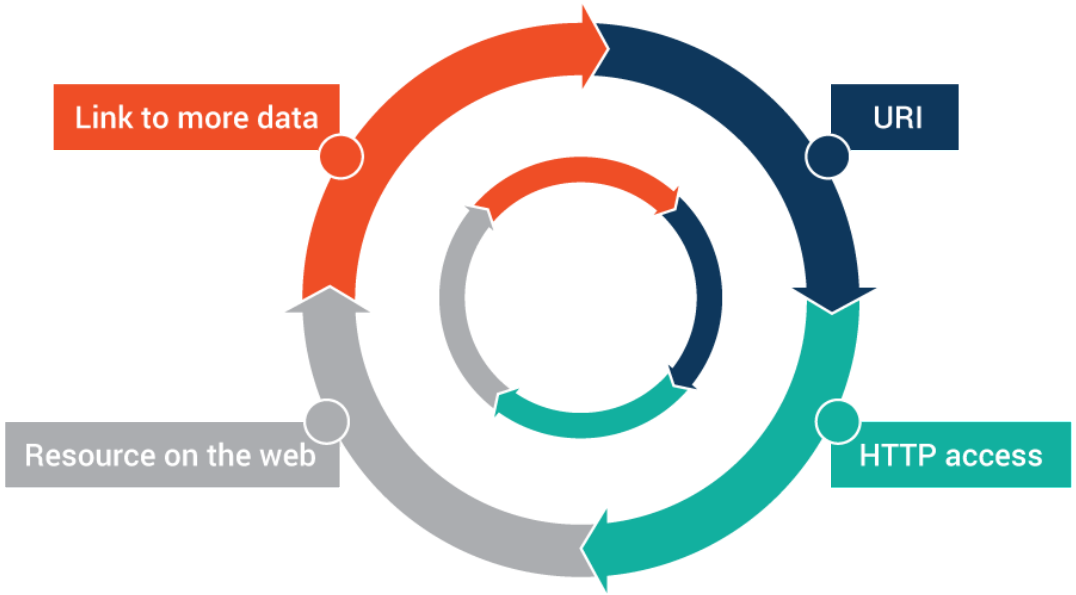
- W3C, extension of the WWW through new standards
- Encoding of semantics with the data
- Emergence of Ontologies
  - FOAF
  - OWL



slide by Basel Shbita, images from https://www.ontotext.com/knowledgehub/fundamentals/what-is-the-semantic-web/

Linked Open Data

- Linked Data + Open Data
  - DBpedia
  - GeoNames
  - Wikidata
- Present day "**Knowledge Graphs**"
  - across vast amount of general importance
  - alive LOD
  - graph-computing techniques and algorithms



slide by Basel Shbita, image from https://www.ontotext.com/knowledgehub/fundamentals/what-is-the-semantic-web/

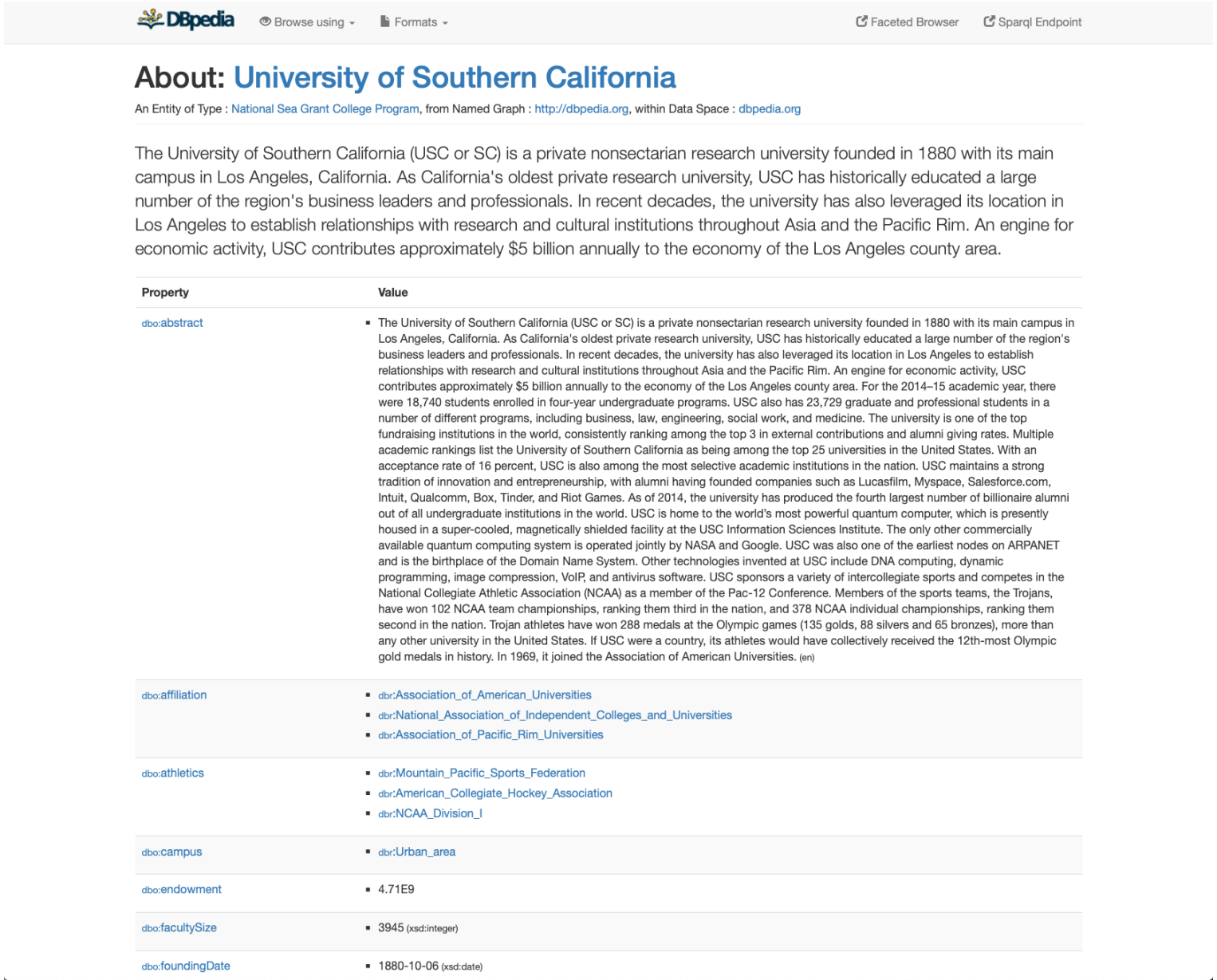
Linked Data Principles

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 RDF information
4. Include RDF statements that link to other URIs so that they can discover more things

http://youtu.be/OM6XIICm_qo

Can USC Have a URI?

http://dbpedia.org/resource/University_of_Southern_California



DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

About: University of Southern California

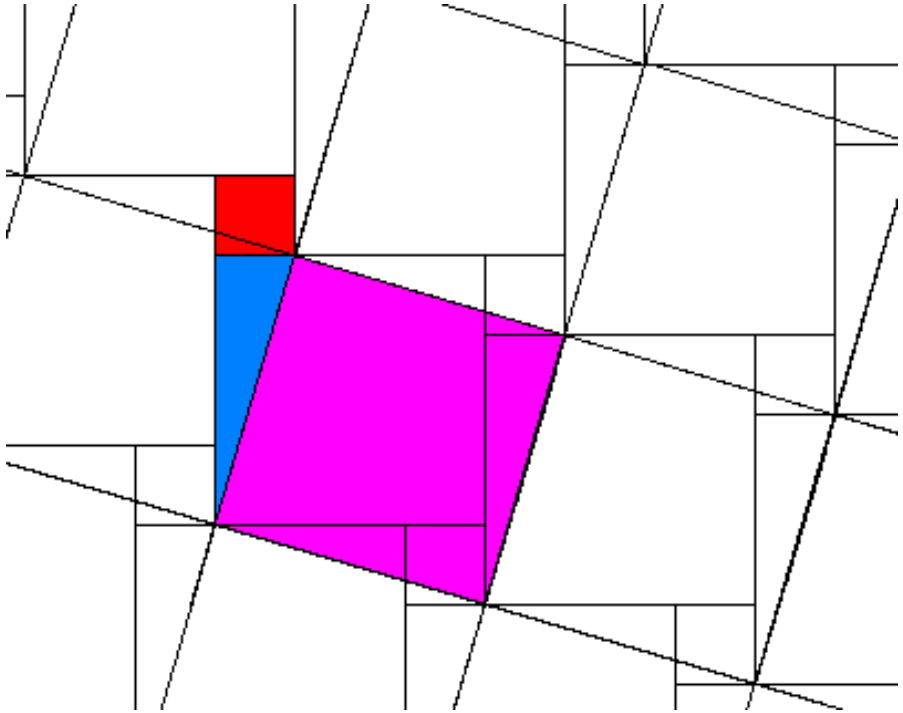
An Entity of Type : National Sea Grant College Program, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

The University of Southern California (USC or SC) is a private nonsectarian research university founded in 1880 with its main campus in Los Angeles, California. As California's oldest private research university, USC has historically educated a large number of the region's business leaders and professionals. In recent decades, the university has also leveraged its location in Los Angeles to establish relationships with research and cultural institutions throughout Asia and the Pacific Rim. An engine for economic activity, USC contributes approximately $5 billion annually to the economy of the Los Angeles county area.

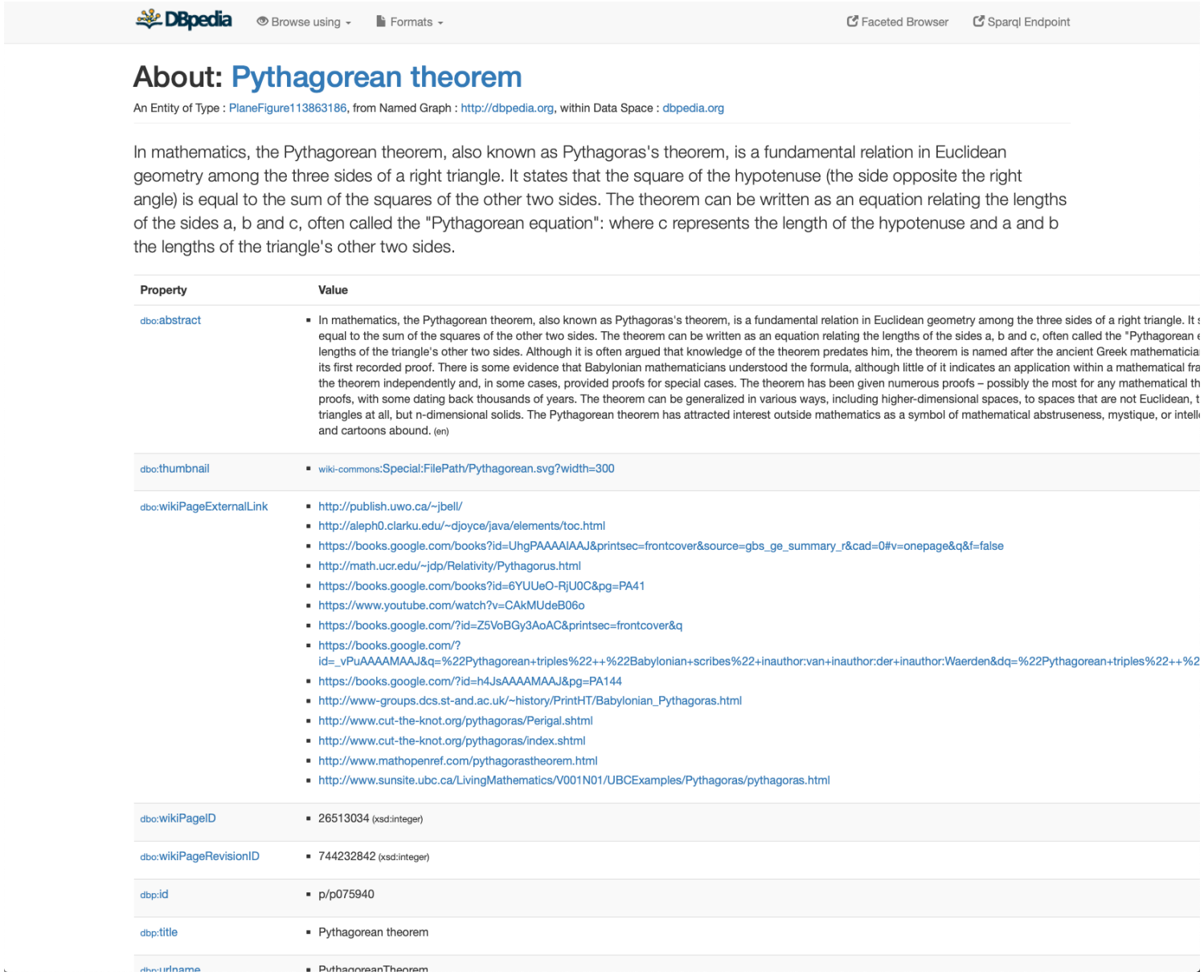
Property	Value
dbo:abstract	▪ The University of Southern California (USC or SC) is a private nonsectarian research university founded in 1880 with its main campus in Los Angeles, California. As California's oldest private research university, USC has historically educated a large number of the region's business leaders and professionals. In recent decades, the university has also leveraged its location in Los Angeles to establish relationships with research and cultural institutions throughout Asia and the Pacific Rim. An engine for economic activity, USC contributes approximately $5 billion annually to the economy of the Los Angeles county area. For the 2014–15 academic year, there were 18,740 students enrolled in four-year undergraduate programs. USC also has 23,729 graduate and professional students in a number of different programs, including business, law, engineering, social work, and medicine. The university is one of the top fundraising institutions in the world, consistently ranking among the top 3 in external contributions and alumni giving rates. Multiple academic rankings list the University of Southern California as being among the top 25 universities in the United States. With an acceptance rate of 16 percent, USC is also among the most selective academic institutions in the nation. USC maintains a strong tradition of innovation and entrepreneurship, with alumni having founded companies such as Lucasfilm, Myspace, Salesforce.com, Intuit, Qualcomm, Box, Tinder, and Riot Games. As of 2014, the university has produced the fourth largest number of billionaire alumni out of all undergraduate institutions in the world. USC is home to the world's most powerful quantum computer, which is presently housed in a super-cooled, magnetically shielded facility at the USC Information Sciences Institute. The only other commercially available quantum computing system is operated jointly by NASA and Google. USC was also one of the earliest nodes on ARPANET and is the birthplace of the Domain Name System. Other technologies invented at USC include DNA computing, dynamic programming, image compression, VoIP, and antivirus software. USC sponsors a variety of intercollegiate sports and competes in the National Collegiate Athletic Association (NCAA) as a member of the Pac-12 Conference. Members of the sports teams, the Trojans, have won 102 NCAA team championships, ranking them third in the nation, and 378 NCAA individual championships, ranking them second in the nation. Trojan athletes have won 288 medals at the Olympic games (135 golds, 88 silvers and 65 bronzes), more than any other university in the United States. If USC were a country, its athletes would have collectively received the 12th-most Olympic gold medals in history. In 1969, it joined the Association of American Universities. (en)
dbo:affiliation	▪ dbr:Association_of_American_Universities ▪ dbr:National_Association_of_Independent_Colleges_and_Universities ▪ dbr:Association_of_Pacific_Rim_Universities
dbo:athletics	▪ dbr:Mountain_Pacific_Sports_Federation ▪ dbr:American_Collegiate_Hockey_Association ▪ dbr:NCAA_Division_I
dbo:campus	▪ dbr:Urban_area
dbo:endowment	▪ 4.71E9
dbo:facultySize	▪ 3945 (xsd:integer)
dbo:foundingDate	▪ 1880-10-06 (xsd:date)

slide by Pedro Szekely

Can the Pythagoras Theorem Have a URI?



http://dbpedia.org/resource/Pythagorean_theorem



DBpedia · Browse using · Formats · Faceted Browser · Sparql Endpoint

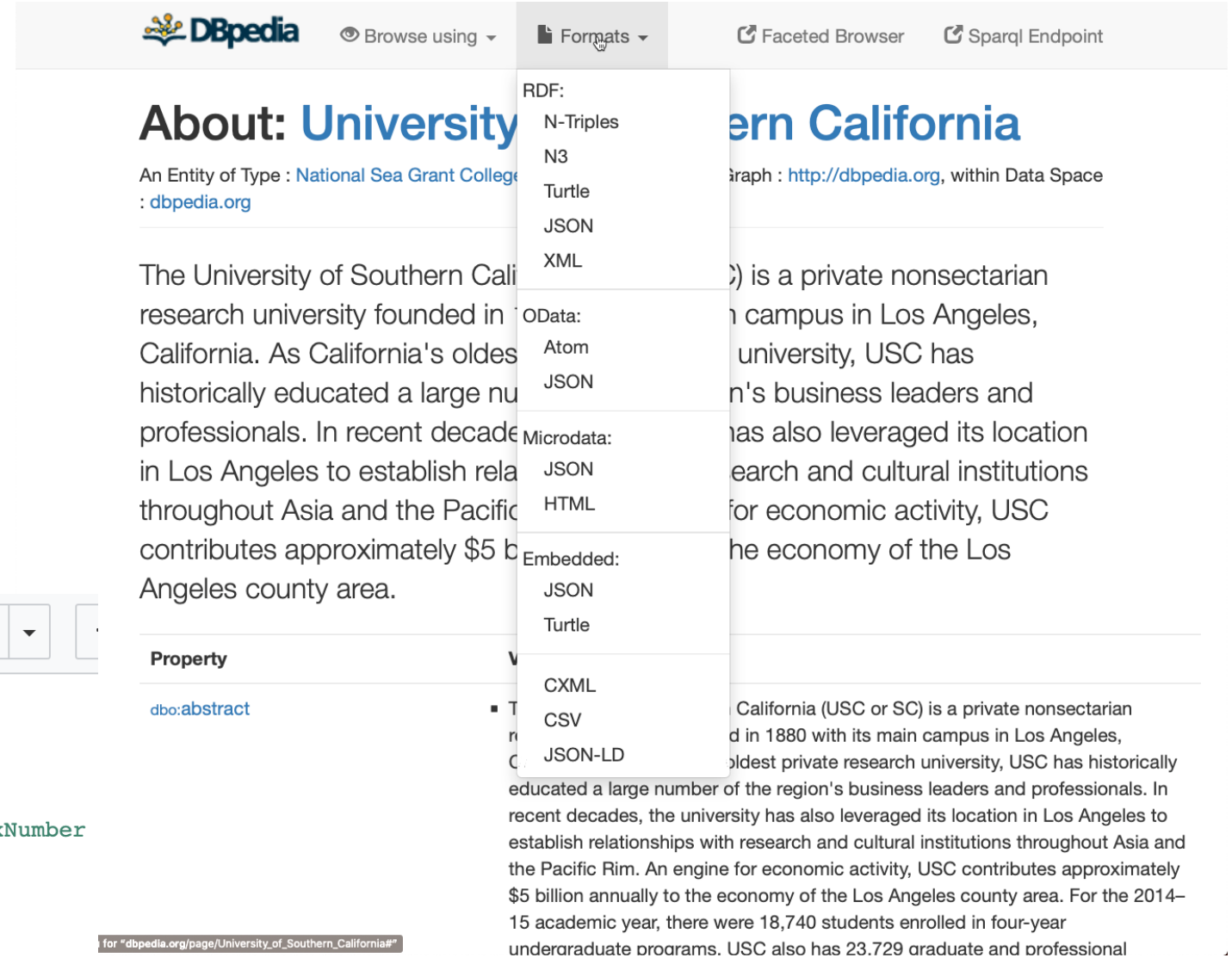
About: Pythagorean theorem

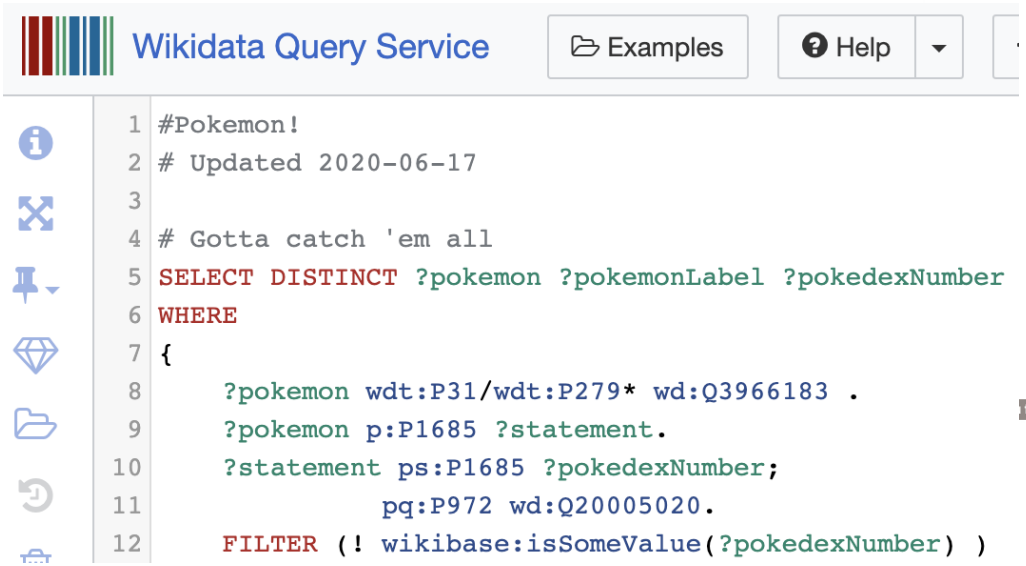
An Entity of Type : PlaneFigure113863186, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

In mathematics, the Pythagorean theorem, also known as Pythagoras's theorem, is a fundamental relation in Euclidean geometry among the three sides of a right triangle. It states that the square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides. The theorem can be written as an equation relating the lengths of the sides a, b and c, often called the "Pythagorean equation": where c represents the length of the hypotenuse and a and b the lengths of the triangle's other two sides.

Property	Value
dbo:abstract	In mathematics, the Pythagorean theorem, also known as Pythagoras's theorem, is a fundamental relation in Euclidean geometry among the three sides of a right triangle. It s equal to the sum of the squares of the other two sides. The theorem can be written as an equation relating the lengths of the sides a, b and c, often called the "Pythagorean e lengths of the triangle's other two sides. Although it is often argued that knowledge of the theorem predates him, the theorem is named after the ancient Greek mathematicia its first recorded proof. There is some evidence that Babylonian mathematicians understood the formula, although little of it indicates an application within a mathematical fra the theorem independently and, in some cases, provided proofs for special cases. The theorem has been given numerous proofs – possibly the most for any mathematical th proofs, with some dating back thousands of years. The theorem can be generalized in various ways, including higher-dimensional spaces, to spaces that are not Euclidean, t triangles at all, but n-dimensional solids. The Pythagorean theorem has attracted interest outside mathematics as a symbol of mathematical abstruseness, mystique, or intell and cartoons abound. (en)
dbo:thumbnail	wiki-commons:Special:FilePath/Pythagorean.svg?width=300
dbo:wikiPageExternalLink	http://publish.uwo.ca/~jbell/ http://aleph0.clarku.edu/~djoyce/java/elements/toc.html https://books.google.com/books?id=UhgPAAAAIAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false http://math.ucr.edu/~jdp/Relativity/Pythagorus.html https://books.google.com/books?id=6YUUeO-RjU0C&pg=PA41 https://www.youtube.com/watch?v=CAkMUdeB06o https://books.google.com/?id=Z5VoBGy3AoAC&printsec=frontcover&q https://books.google.com/?id=_vPuAAAAMAAJ&q=%22Pythagorean+triples%22++%22Babylonian+scribes%22+inauthor:van+inauthor:der+inauthor:Waerden&dq=%22Pythagorean+triples%22++%2 https://books.google.com/?id=h4JsAAAAMAAJ&pg=PA144 http://www-groups.dcs.st-and.ac.uk/~history/PrintHT/Babylonian_Pythagoras.html http://www.cut-the-knot.org/pythagoras/Perigal.shtml http://www.cut-the-knot.org/pythagoras/index.shtml http://www.mathopenref.com/pythagorastheorem.html http://www.sunsite.ubc.ca/LivingMathematics/V001N01/UBCExamples/Pythagoras/pythagoras.html
dbo:wikiPageID	26513034 (xsd:integer)
dbo:wikiPageRevisionID	744232842 (xsd:integer)
dbp:id	p/p075940
dbp:title	Pythagorean theorem
dbp:urlname	PythagoreanTheorem

slide by Pedro Szekely





USC sponsors a variety of intercollegiate sports and competes in the National Collegiate Athletic Association (NCAA) as a member of the Pac-12 Conference. Members of the sports teams, the T have won 102 NCAA team championships, ranking them third in the nation, and 378 NCAA indiv championships, ranking them second in the nation. Trojan athletes have won 288 medals at the (games (135 golds, 88 silvers and 65 bronzes), more than any other university in the United State were a country, its athletes would have collectively received the 12th-most Olympic gold medals history. In 1969, it joined the Association of American Universities. (en)

dbo:affiliation	▪ dbr:Association_of_American_Universities ▪ dbr:National_Association_of_Independent_Colleges_and_Universities ▪ dbr:Association_of_Pacific_Rim_Universities
dbo:athletics	▪ dbr:Mountain_Pacific_Sports_Federation ▪ dbr:American_Collegiate_Hockey_Association ▪ dbr:NCAA_Division_I
dbo:campus	▪ dbr:Urban_area
dbo:endowment	▪ 4.71E9
dbo:facultySize	▪ 3945 (xsd:integer)
dbo:foundingDate	▪ 1880-10-06 (xsd:date)

Now we know what linked data is

What can go wrong?

Different URIs For the Same Thing

- https://www.wikidata.org/wiki/Q36107
- http://dbpedia.org/resource/Muhammad_Ali
- https://yago-knowledge.org/resource/Muhammad_Ali
- http://data.nytimes.com/N13611972026987463463

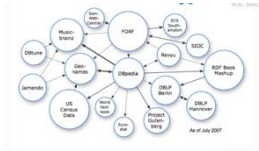
Linked Data Challenges

- Different URIs for the same thing
  - … makes it harder to link the data
- Timeliness
  - … not up to date
- Provenance
  - … not only a linked data problem
- Tools
  - … slow performance compared to traditional data
  - … search engines not yet mature
  - … many RDF formats, not supported by all tools

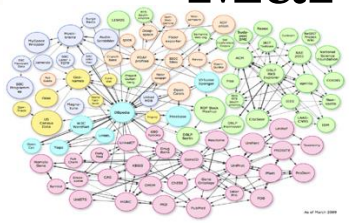
slide by Pedro Szekely

Getting Bigger Every Day

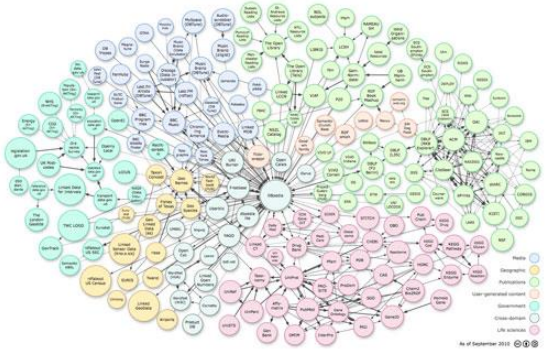
July 2007



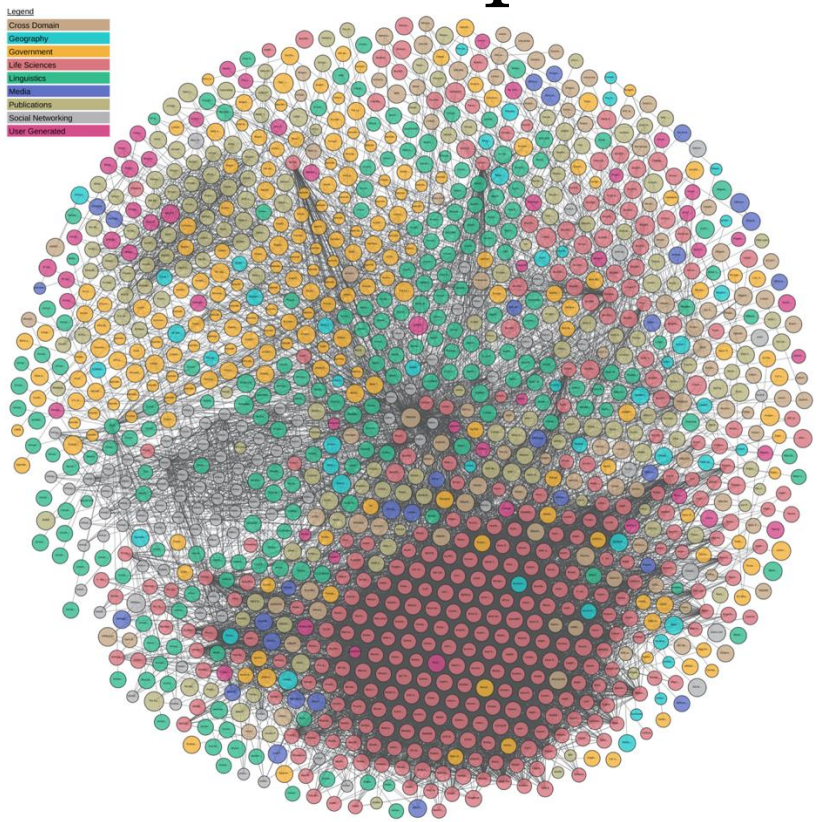
March 2009



September 2010



April 2020



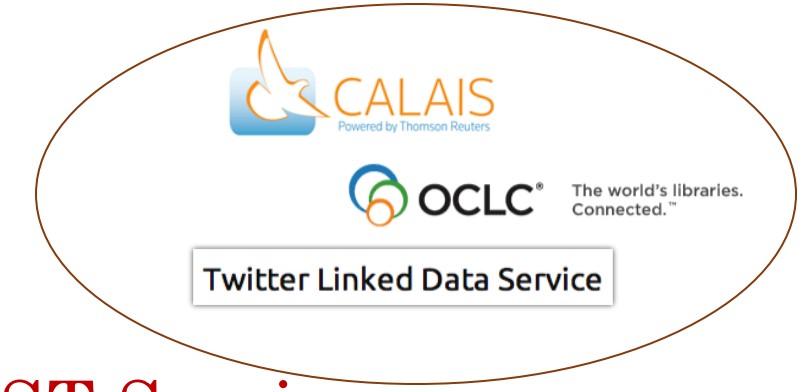
slide by Pedro Szekely

Working with Linked Data

RDF Dump

SPARQL Endpoint

URI Derefencing



REST Service

slide by Pedro Szekely

Is your Linked Open Data 5 Star?

Available on the web (whatever format) *but with an open licence, to be Open Data*

Available as machine-readable structured data (e.g. excel instead of image scan of a table)

as (2) plus non-proprietary format (e.g. CSV instead of excel)

All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

All the above, plus: Link your data to other people's data to provide context

Best Practices for Data on the Web

https://www.w3.org/TR/dwbp

Best Practice 1: Provide metadata

Best Practice 2: Provide descriptive metadata

Best Practice 3: Provide structural metadata

Best Practice 4: Provide data license information

Best Practice 5: Provide data provenance information

Best Practice 6: Provide data quality information

Best Practice 7: Provide a version indicator

Best Practice 8: Provide version history

Best Practice 9: Use persistent URIs as identifiers of datasets

Best Practice 10: Use persistent URIs as identifiers within datasets

Best Practice 11: Assign URIs to dataset versions and series

Best Practice 12: Use machine-readable standardized data formats

Best Practice 13: Use locale-neutral data representations

Best Practice 14: Provide data in multiple formats

Best Practice 15: Reuse vocabularies, preferably standardized ones

Best Practice 16: Choose the right formalization level

Best Practice 17: Provide bulk download

Best Practice 18: Provide Subsets for Large Datasets

Best Practice 19: Use content negotiation for serving data available in multiple formats

Best Practice 20: Provide real-time access

Best Practice 21: Provide data up to date

Best Practice 22: Provide an explanation for data that is not available

Best Practice 23: Make data available through an API

Best Practice 24: Use Web Standards as the foundation of APIs

Best Practice 25: Provide complete documentation for your API

Best Practice 26: Avoid Breaking Changes to Your API

Best Practice 27: Preserve identifiers

Best Practice 28: Assess dataset coverage

Best Practice 29: Gather feedback from data consumers

Best Practice 30: Make feedback available

Best Practice 31: Enrich data by generating new data

Best Practice 32: Provide Complementary Presentations

Best Practice 33: Provide Feedback to the Original Publisher

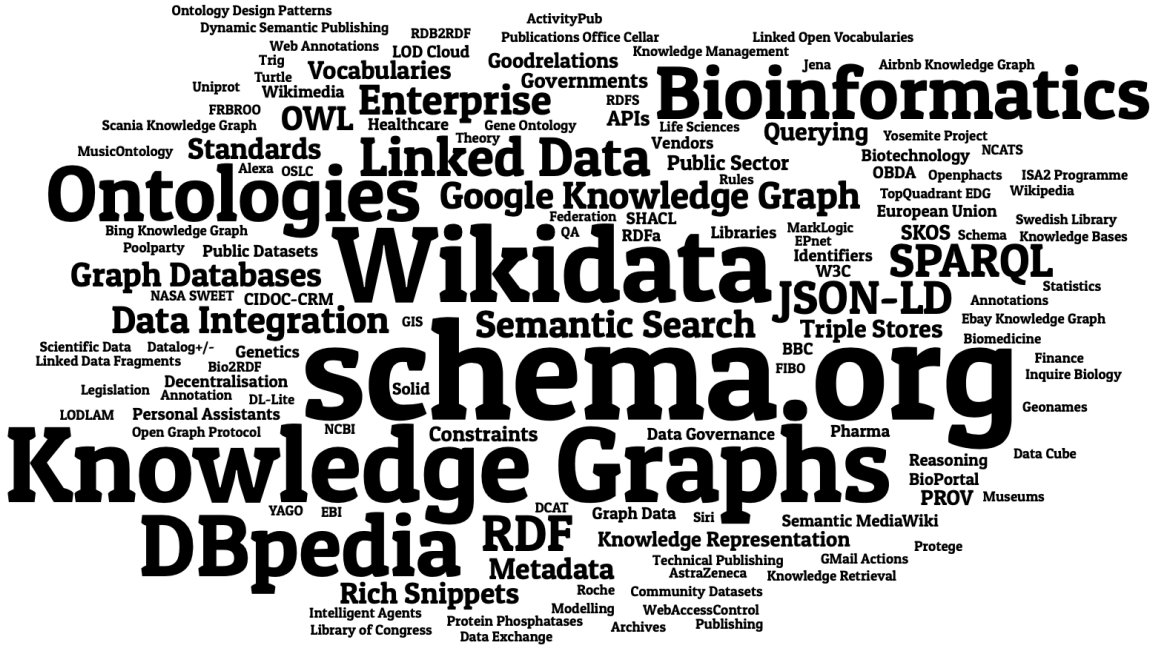
Best Practice 34: Follow Licensing Terms

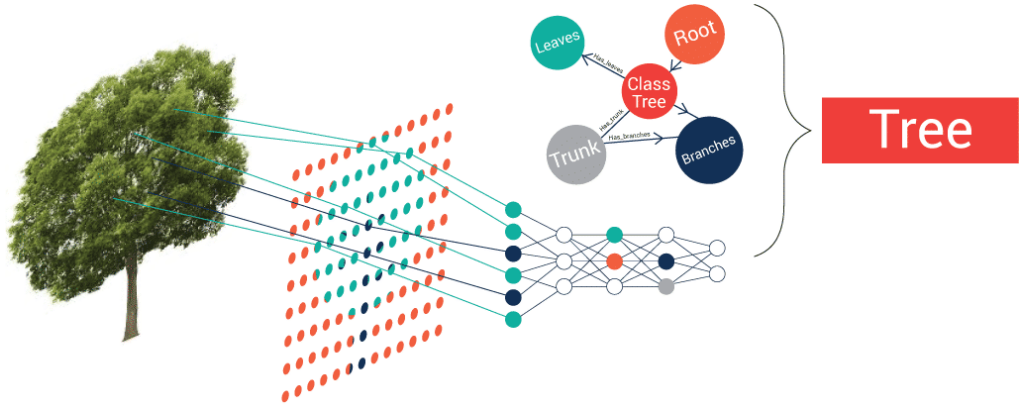
Best Practice 35: Cite the Original Publication

slide by Basel Shbita

What's next?

- Players
  - Academia
  - Industry
- Criticisms
  - lack of usable tools
  - lack of incentives
  - lack of robustness for unreliable publishers
  - overly verbose standards
- Vision
  - bridge the gap, in particular, to developers
  - inject Semantic Web interoperability into already broadly-adopted technologies





slide by Basel Shbita, images from https://www.ontotext.com/knowledgehub/fundamentals/what-is-the-semantic-web/