Libraries at the Crossroads

Resolving Identities

EMEA
REGIONAL COUNCIL MEETING
BERLIN
21-22 FEB. 2017

OCLC®

#EMEARC17

EMEA

REGIONAL COUNCIL MEETING

BERLIN

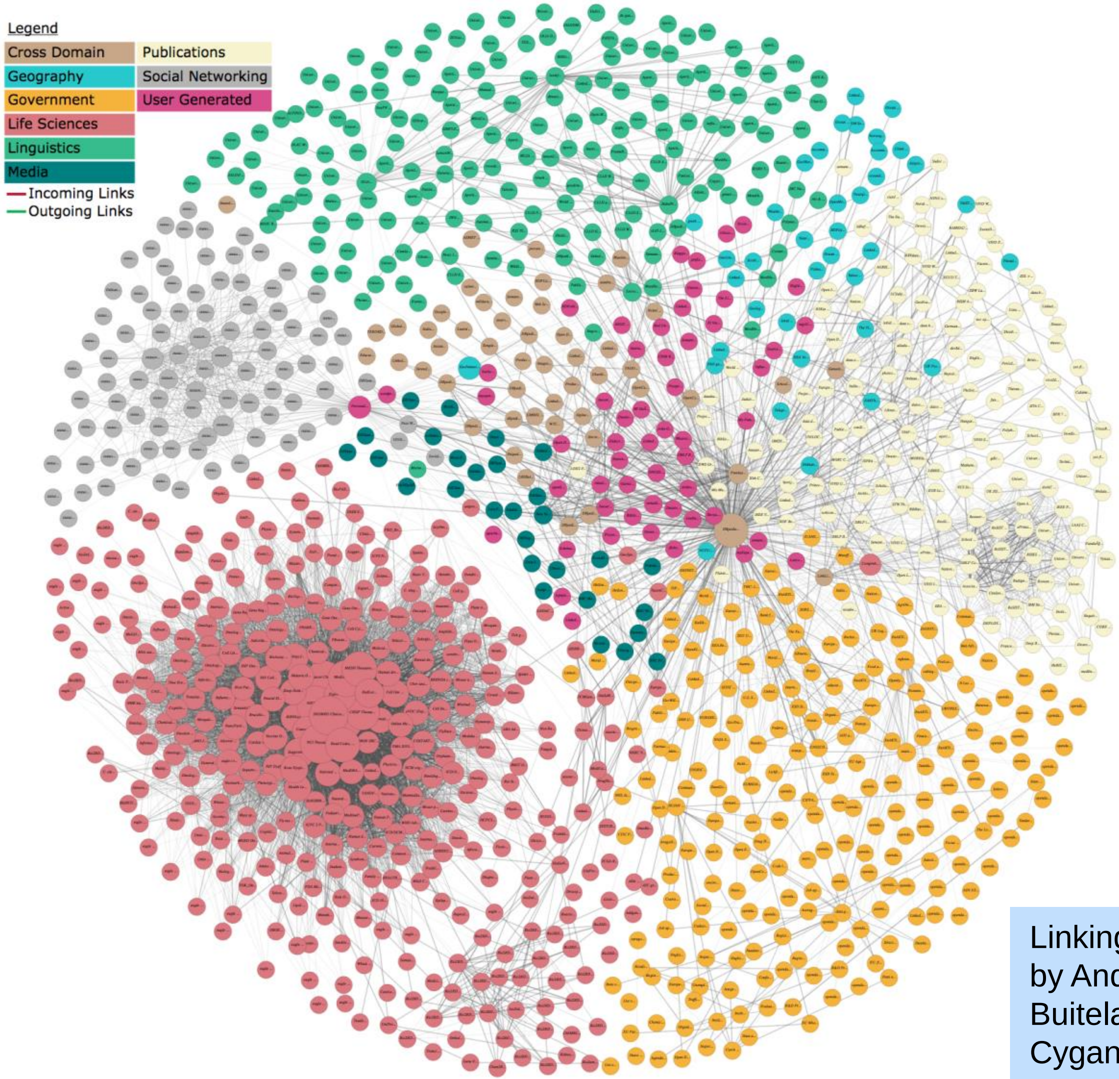
21-22 FEB. 2017

OCLC®

Linking entities via semantic indexing

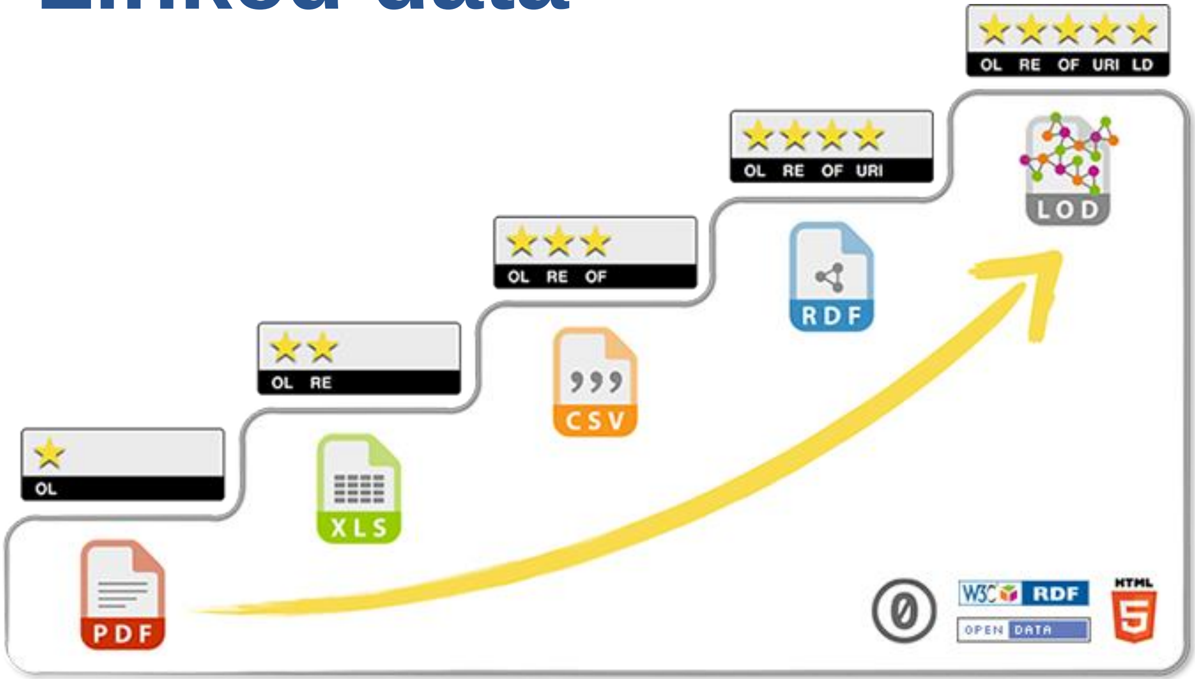
Shenghui Wang, Rob Koopman

#EMEARC17



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

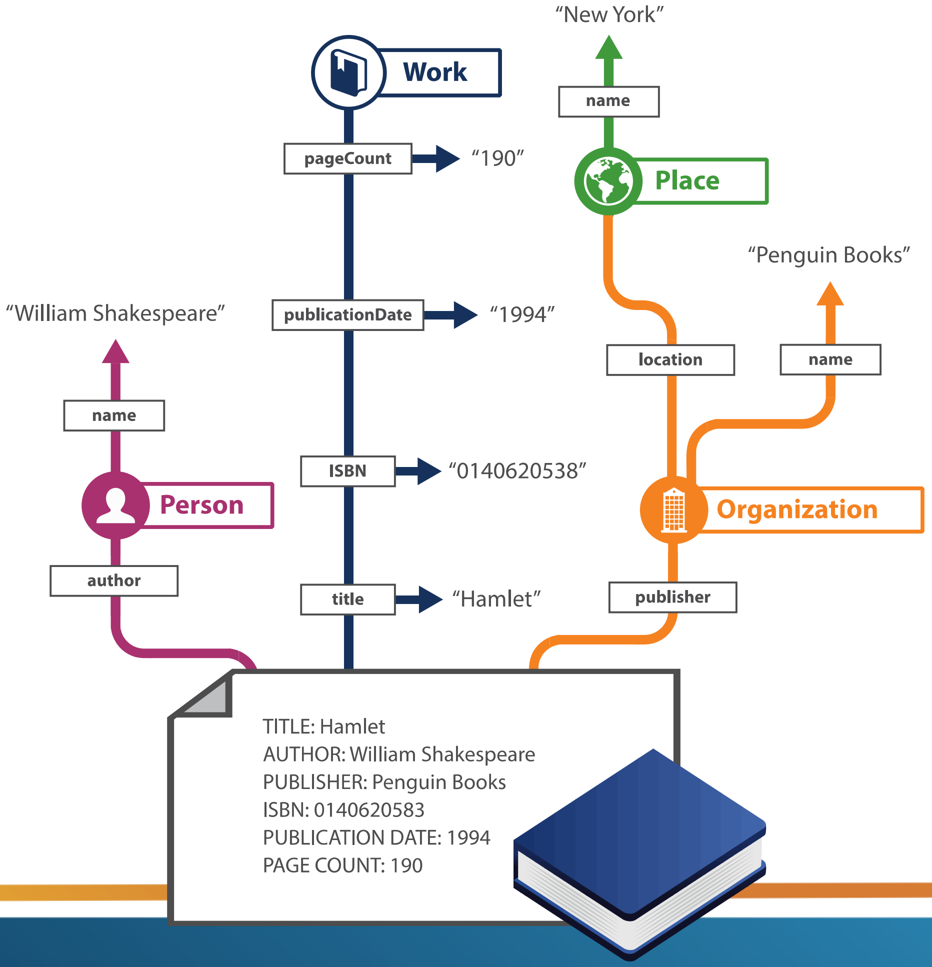
Linked data



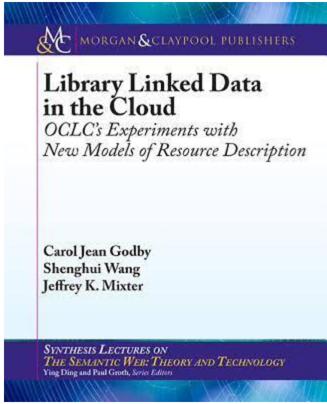
http://5stardata.info/en/

The Semantic Web isn't just about putting data on the web. It is about **making links**, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Tim Berners-Lee



TITLE: Hamlet
AUTHOR: William Shakespeare
PUBLISHER: Penguin Books
ISBN: 0140620583
PUBLICATION DATE: 1994
PAGE COUNT: 190

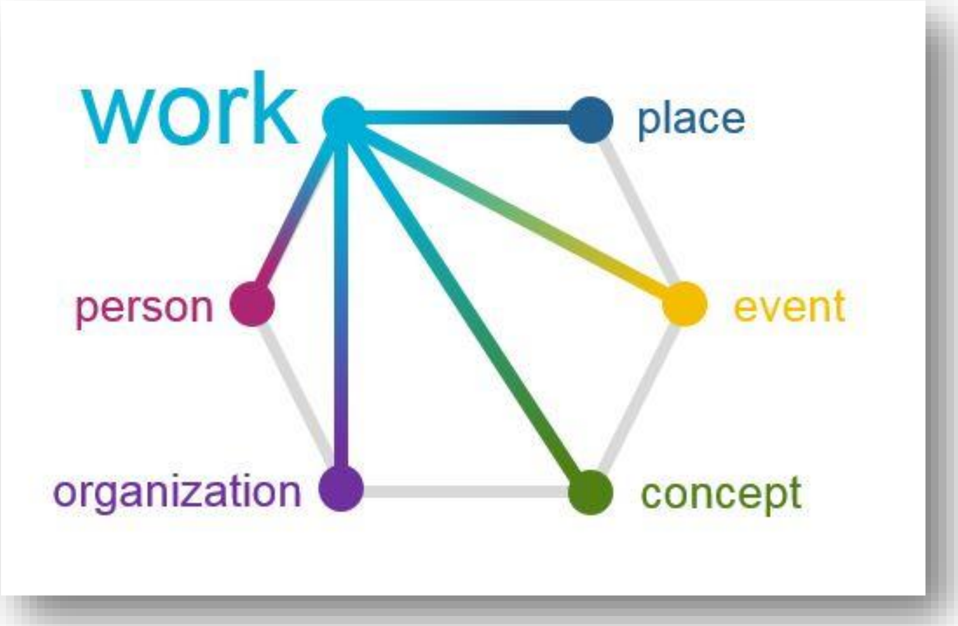

MORGAN&CLAYPOOL PUBLISHERS

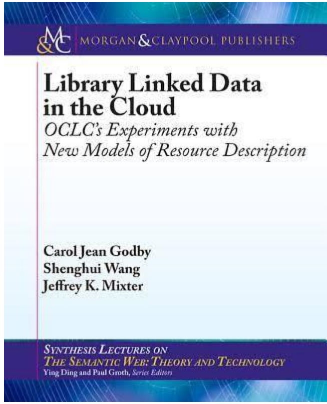
Library Linked Data in the Cloud

OCLC's Experiments with New Models of Resource Description

Carol Jean Godby
Shenghui Wang
Jeffrey K. Mixter

Synthesis Lectures on The Semantic Web: Theory and Technology
Ying Ding and Paul Groth, *Series Editors*



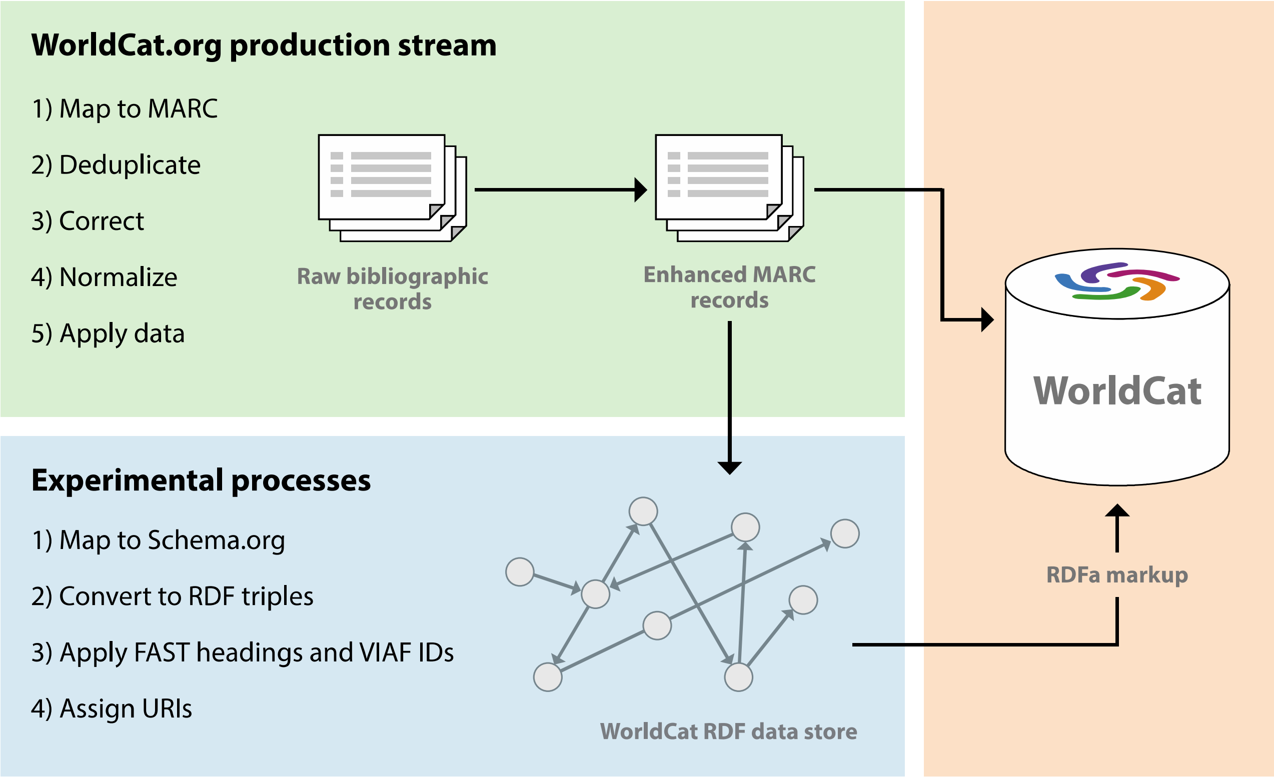

MORGAN & CLAYPOOL PUBLISHERS

Library Linked Data in the Cloud

OCLC's Experiments with New Models of Resource Description

Carol Jean Godby
Shenghui Wang
Jeffrey K. Mixter

Synthesis Lectures on The Semantic Web: Theory and Technology
Ying Ding and Paul Groth, *Series Editors*



WorldCat.org production stream

1) Map to MARC

2) Deduplicate

3) Correct

4) Normalize

5) Apply data

Raw bibliographic records

Enhanced MARC records

WorldCat

RDFa markup

Experimental processes

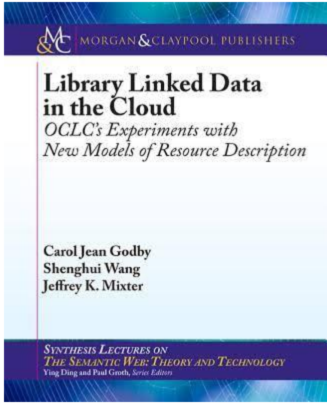
1) Map to Schema.org

2) Convert to RDF triples

3) Apply FAST headings and VIAF IDs

4) Assign URIs

WorldCat RDF data store



MORGAN & CLAYPOOL PUBLISHERS

Library Linked Data in the Cloud

OCLC's Experiments with New Models of Resource Description

Carol Jean Godby
Shenghui Wang
Jeffrey K. Mixter

SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY

Ying Ding and Paul Groth, *Series Editors*

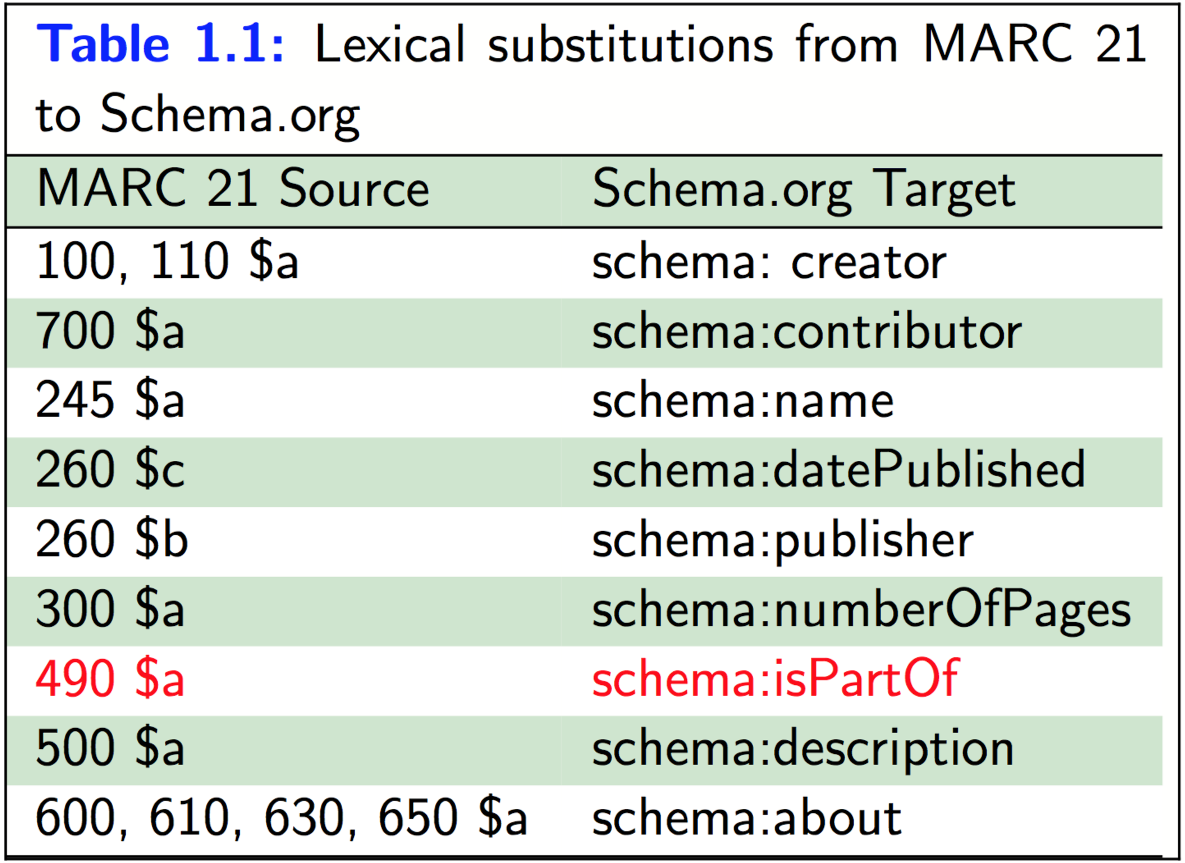
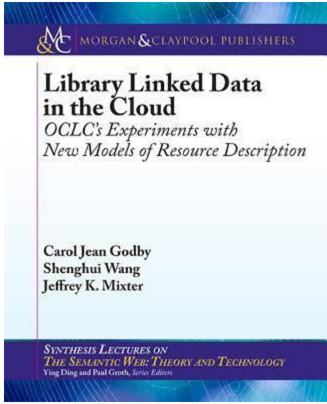


Table 1.1: Lexical substitutions from MARC 21 to Schema.org

MARC 21 Source	Schema.org Target
100, 110 $a	schema: creator
700 $a	schema:contributor
245 $a	schema:name
260 $c	schema:datePublished
260 $b	schema:publisher
300 $a	schema:numberOfPages
490 $a	schema:isPartOf
500 $a	schema:description
600, 610, 630, 650 $a	schema:about



MORGAN & CLAYPOOL PUBLISHERS

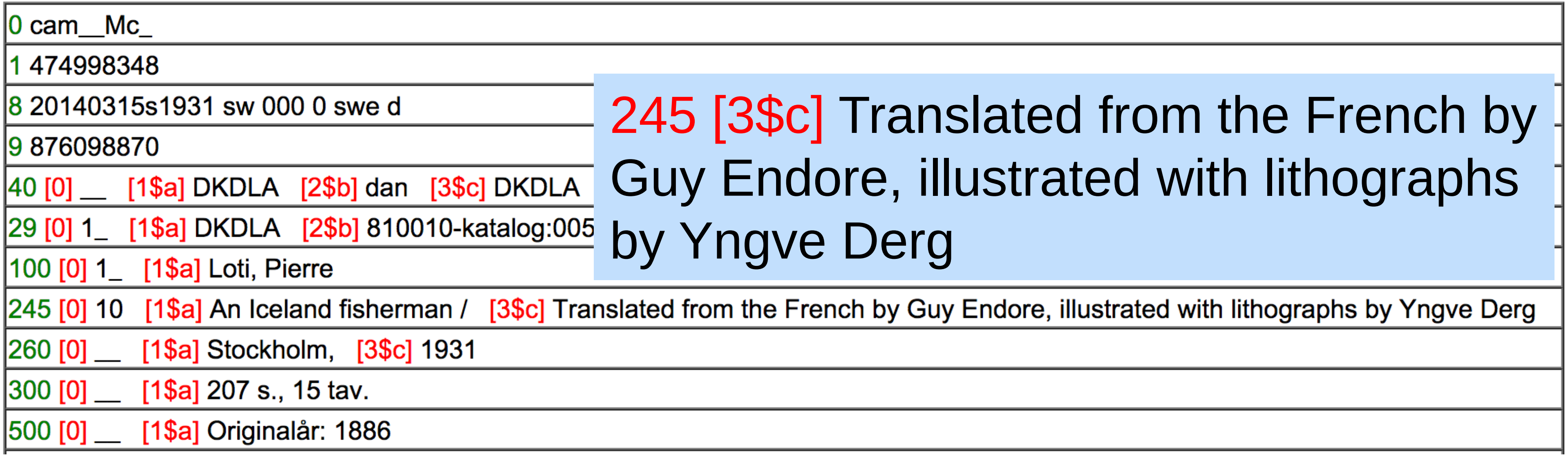
Library Linked Data in the Cloud

OCLC's Experiments with New Models of Resource Description

Carol Jean Godby
Shenghui Wang
Jeffrey K. Mixter

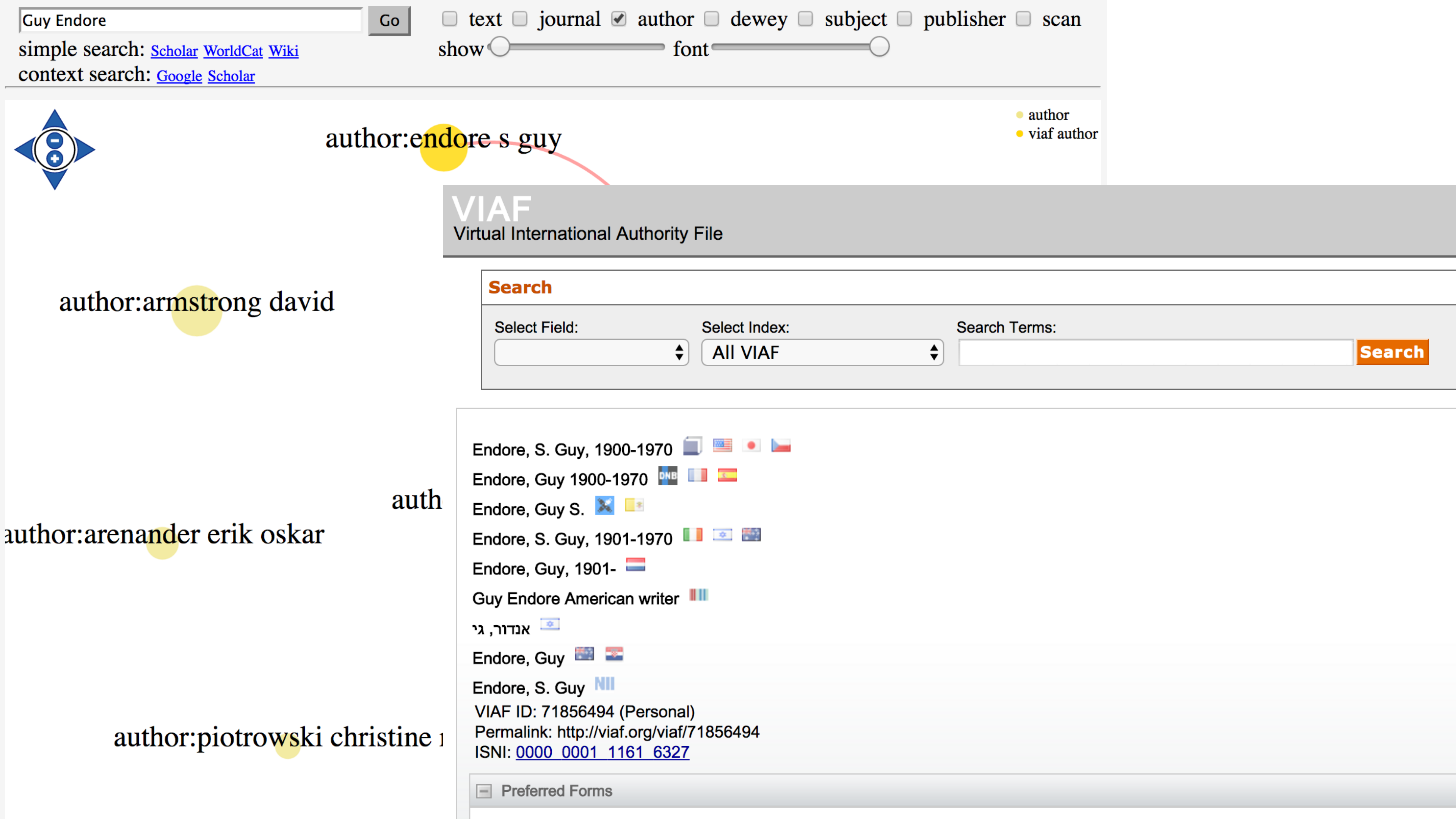
SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY
Ying Ding and Paul Groth, *Series Editors*

Information locked in free-text



```
0 cam__Mc_
1 474998348
8 20140315s1931 sw 000 0 swe d
9 876098870
40 [0] __   [1$a] DKDLA   [2$b] dan   [3$c] DKDLA
29 [0] 1_   [1$a] DKDLA   [2$b] 810010-katalog:005
100 [0] 1_   [1$a] Loti, Pierre
245 [0] 10   [1$a] An Iceland fisherman /   [3$c] Translated from the French by Guy Endore, illustrated with lithographs by Yngve Derg
260 [0] __   [1$a] Stockholm,   [3$c] 1931
300 [0] __   [1$a] 207 s., 15 tav.
500 [0] __   [1$a] Originalår: 1886
```

245 [3$c] Translated from the French by Guy Endore, illustrated with lithographs by Yngve Derg



Guy Endore Go □ text □ journal ☑ author □ dewey □ subject □ publisher □ scan

simple search: Scholar WorldCat Wiki show font

context search: Google Scholar

VIAF
Virtual International Authority File

Search

Select Field: Select Index: All VIAF Search Terms: Search

Endore, S. Guy, 1900-1970

Endore, Guy 1900-1970

Endore, Guy S.

Endore, S. Guy, 1901-1970

Endore, Guy, 1901-

Guy Endore American writer

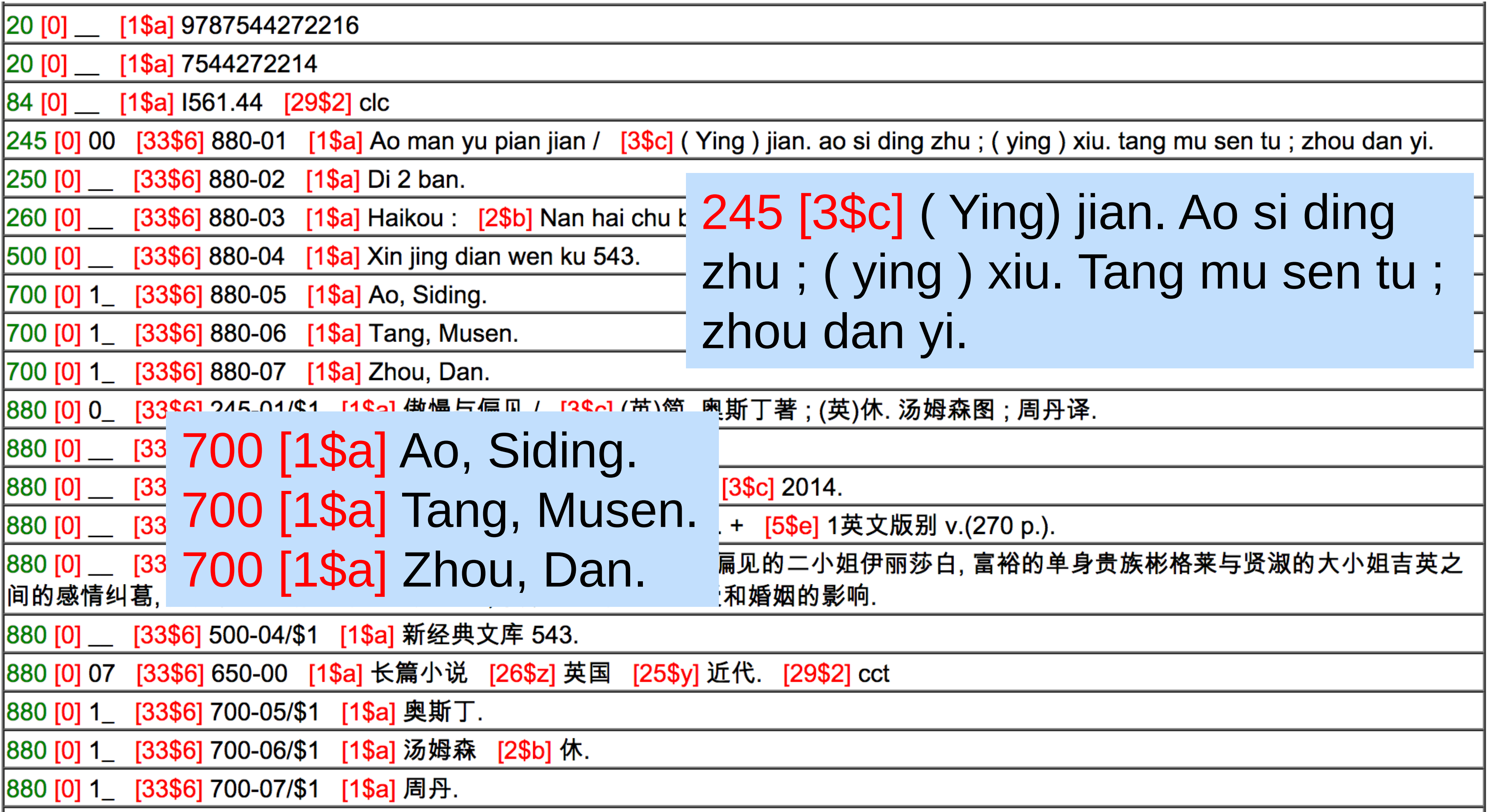
אנדור, גי

Endore, Guy

Endore, S. Guy

VIAF ID: 71856494 (Personal)
Permalink: http://viaf.org/viaf/71856494
ISNI: 0000 0001 1161 6327

Preferred Forms



20 [0] __ [1$a] 9787544272216

20 [0] __ [1$a] 7544272214

84 [0] __ [1$a] I561.44 [29$2] clc

245 [0] 00 [33$6] 880-01 [1$a] Ao man yu pian jian / [3$c] (Ying) jian. ao si ding zhu ; (ying) xiu. tang mu sen tu ; zhou dan yi.

250 [0] __ [33$6] 880-02 [1$a] Di 2 ban.

260 [0] __ [33$6] 880-03 [1$a] Haikou : [2$b] Nan hai chu b

500 [0] __ [33$6] 880-04 [1$a] Xin jing dian wen ku 543.

700 [0] 1_ [33$6] 880-05 [1$a] Ao, Siding.

700 [0] 1_ [33$6] 880-06 [1$a] Tang, Musen.

700 [0] 1_ [33$6] 880-07 [1$a] Zhou, Dan.

880 [0] 0_ [33$6] 245-01/$1 [1$a] 傲慢与偏见 / [3$c] (英)简. 奥斯丁著 ; (英)休. 汤姆森图 ; 周丹译.

880 [0] __ [33

880 [0] __ [33 [3$c] 2014.

880 [0] __ [33 . + [5$e] 1英文版别 v.(270 p.).

880 [0] __ [33 扁见的二小姐伊丽莎白, 富裕的单身贵族彬格莱与贤淑的大小姐吉英之间的感情纠葛, 和婚姻的影响.

880 [0] __ [33$6] 500-04/$1 [1$a] 新经典文库 543.

880 [0] 07 [33$6] 650-00 [1$a] 长篇小说 [26$z] 英国 [25$y] 近代. [29$2] cct

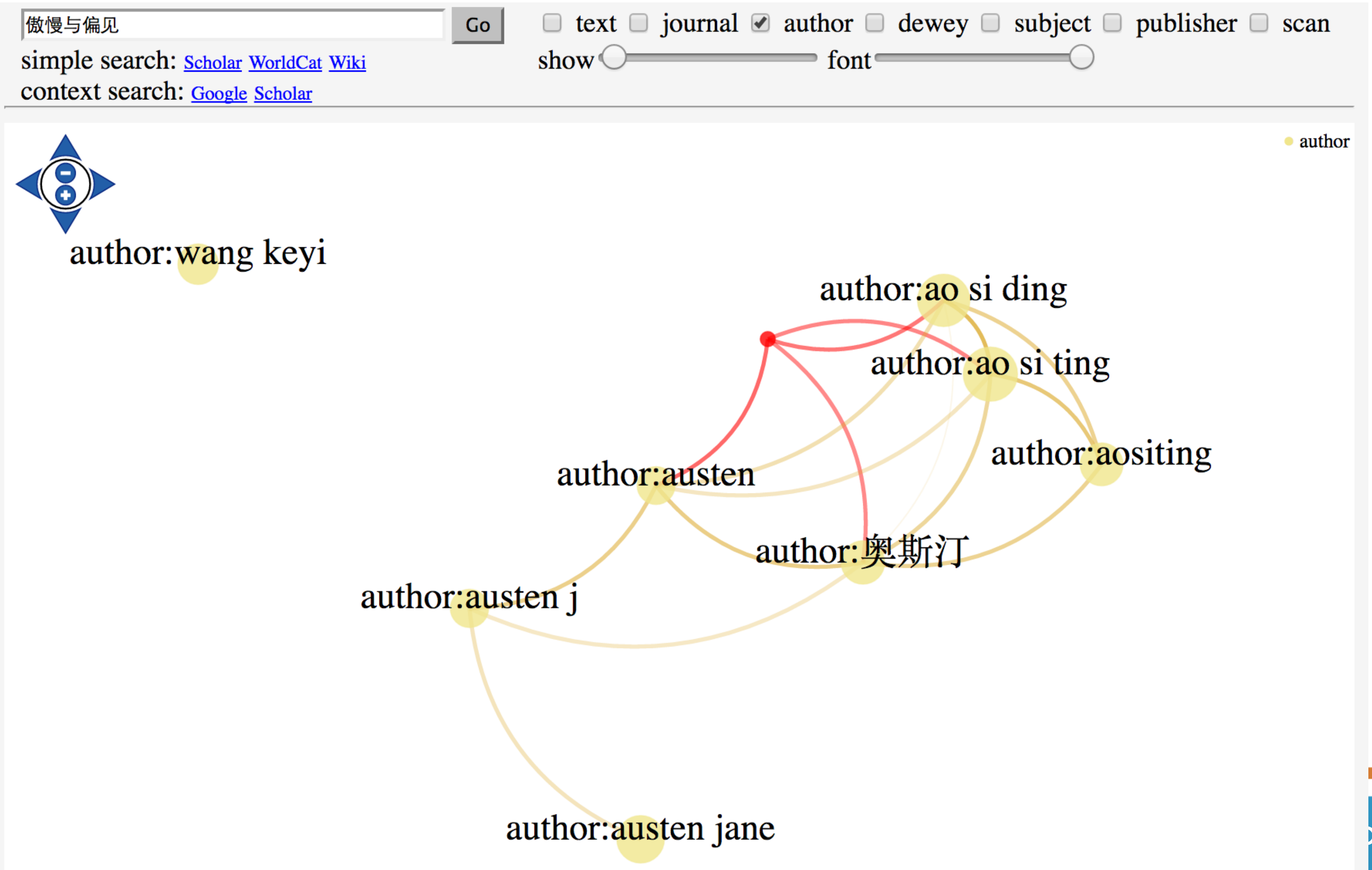
880 [0] 1_ [33$6] 700-05/$1 [1$a] 奥斯丁.

880 [0] 1_ [33$6] 700-06/$1 [1$a] 汤姆森 [2$b] 休.

880 [0] 1_ [33$6] 700-07/$1 [1$a] 周丹.

245 [3$c] (Ying) jian. Ao si ding zhu ; (ying) xiu. Tang mu sen tu ; zhou dan yi.

700 [1$a] Ao, Siding.
700 [1$a] Tang, Musen.
700 [1$a] Zhou, Dan.



傲慢与偏见 Go

☐ text ☐ journal ☑ author ☐ dewey ☐ subject ☐ publisher ☐ scan

simple search: Scholar WorldCat Wiki

show font

context search: Google Scholar

More links could be recovered

- If we have enough (good) data
- If we have effective and scalable algorithms
- If we have patience

Linking entities via semantic indexing

An example by Stefan Evert: what’s the meaning of *bardiwac*?

- He handed her her glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia’s sunshine.
- I dined on bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

⇒ ‘bardiwac’ is a heavy red alcoholic beverage made from grapes

Linking entities via semantic indexing

- *Statistical Semantics* [furnas1983,weaver1955] based on the assumption of “a word is characterized by the company it keeps” [firth1957]
- *Distributional Hypothesis* [harris1954, sahlgren2008]: words that occur in similar contexts tend to have similar meanings

Let’s embed entities in a vector space

- Discrete encoding does not help to automatically process the underlying semantics
- Entities (words) are represented in a continuous vector space where semantically similar words are mapped to nearby points (‘are embedding nearby each other’)
- A desirable property: computable similarity

Word embedding techniques

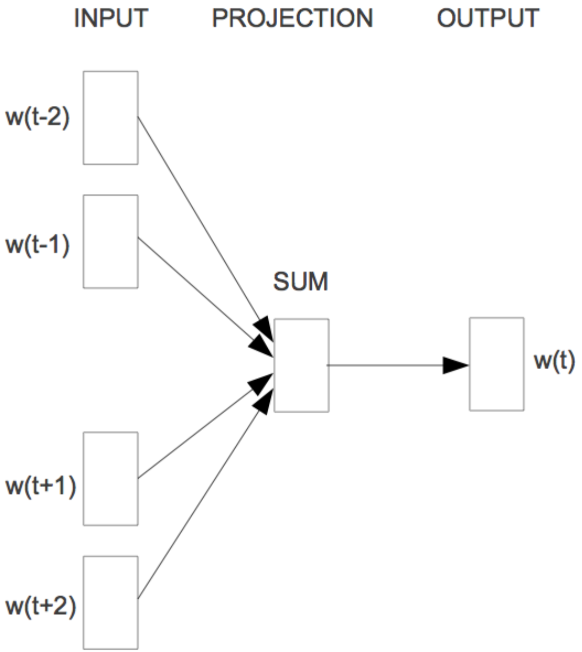
Two main categories of approaches:

- Global co-occurrence count-based method, such as Latent Semantic Analysis
- Local context predictive methods, such as neural probabilistic language models

Word2Vec: Continuous Bag of Words model

- Scan text in large corpus with a window
- The model predicts the current word given the context

the	cat	chills	on	a	mat
w(t-2)	w(t-1)	w(t)	w(t+1)	w(t+2)	

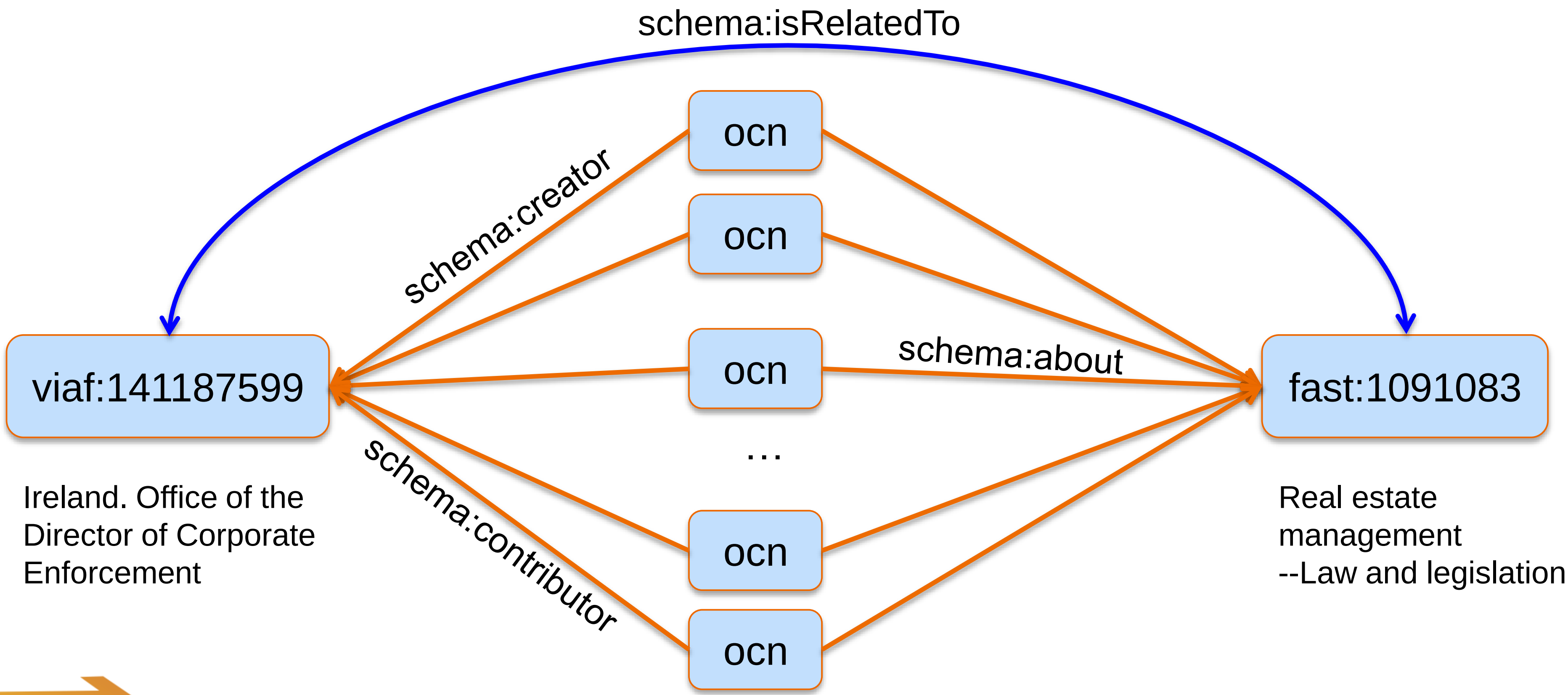


Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

When an entity becomes a vector

- Similarity or relatedness can be computed automatically
  - Cosine similarity
- Such similarity/relatedness can be used to link an entity to its most related entities via **schema:isRelatedTo**
- Such links can be complementary to existing triples

Cosine similarity (viaf:141187599, fast:1091083) = 0.2

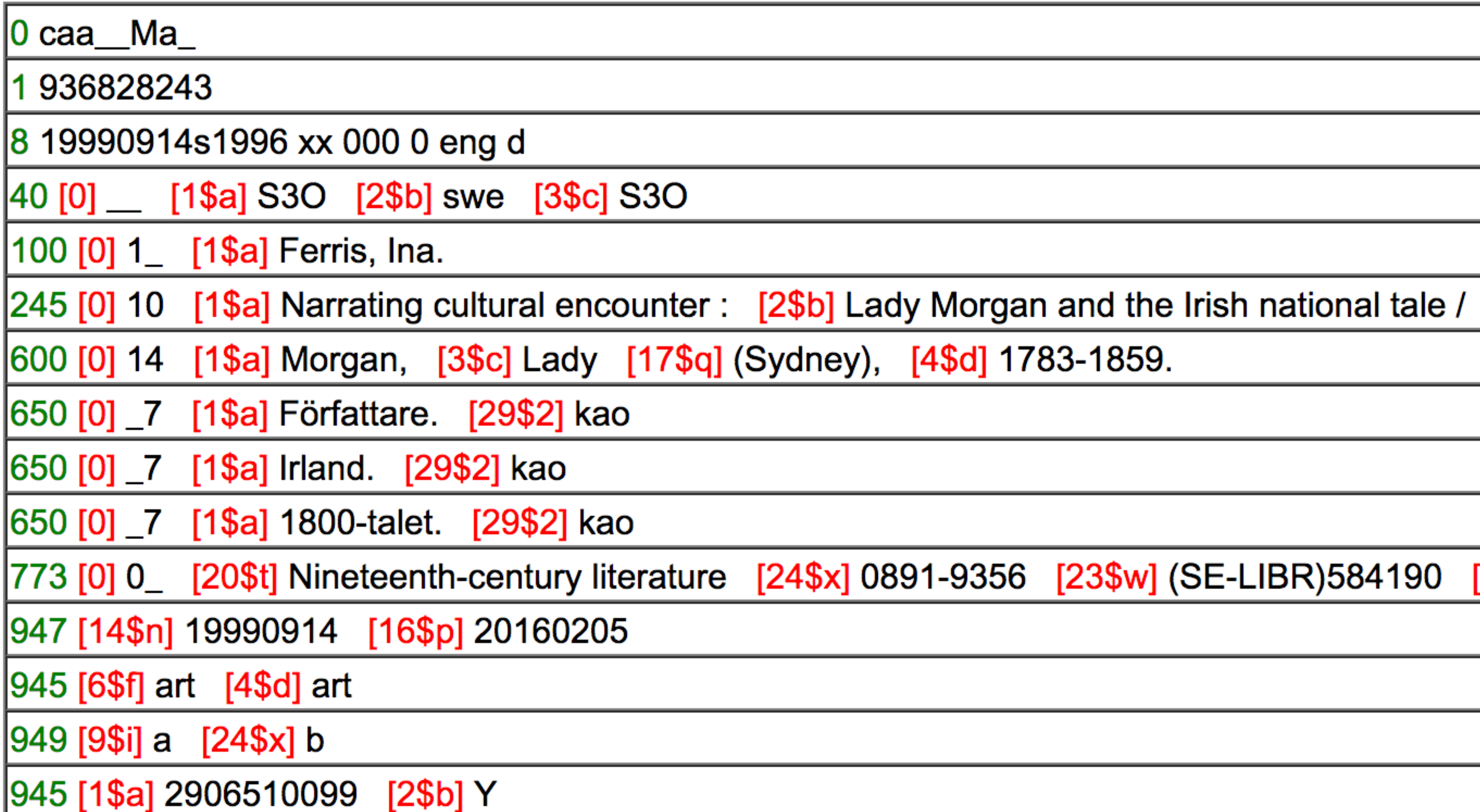


Ireland. Office of the Director of Corporate Enforcement

Real estate management
--Law and legislation

Complementary links

- viaf:41915577 (Ferris, Ina) is mostly related to
  - viaf:68928644 (Engels, Friedrich)
  - fast:1033899 (Nationalism in literature)
  - viaf:57397450 (Smith, Goldwin)
  - viaf:27899280 (Duffy, Charles Gavan)
  - viaf:49228757 (Marx, Karl)
  - viaf:47158897 (McCarthy, Michael John Fitzgerald)
  - viaf:56715842 (Kinealy, Christine)
  - viaf:107533016 (Sydney, Lady Morgan Irish novelist)



```
0 caa__Ma_
1 936828243
8 19990914s1996 xx 000 0 eng d
40 [0] __   [1$a] S3O   [2$b] swe   [3$c] S3O
100 [0] 1_   [1$a] Ferris, Ina.
245 [0] 10   [1$a] Narrating cultural encounter :   [2$b] Lady Morgan and the Irish national tale /
600 [0] 14   [1$a] Morgan,   [3$c] Lady   [17$q] (Sydney),   [4$d] 1783-1859.
650 [0] _7   [1$a] Författare.   [29$2] kao
650 [0] _7   [1$a] Irland.   [29$2] kao
650 [0] _7   [1$a] 1800-talet.   [29$2] kao
773 [0] 0_   [20$t] Nineteenth-century literature   [24$x] 0891-9356   [23$w] (SE-LIBR)584190   [
947 [14$n] 19990914   [16$p] 20160205
945 [6$f] art   [4$d] art
949 [9$i] a   [24$x] b
945 [1$a] 2906510099   [2$b] Y
```

Word embedding techniques

- *Ariadne* (OCLC): based on Random Projection of the global co-occurrence matrix
- *Word2Vec* (Google): shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words
- *GloVe* (Stanford): a global log-bilinear regression model to learn word vectors based on the ratio of the co-occurrence probabilities of two words.

Different model, different embedding

knee	Word2Vec	ankle, hip, elbow, knees, shoulder, patellofemoral, joint, wrist, tka, patellar
	GloVe	ankle, hip, joint, knees, arthroplasty, osteoarthritis, elbow, flexion, cruciate, joints
	Ariadne	knees, knee joint, contralateral knee, tibiofemoral, knee pain, knee motion, medial compartment, lateral compartment, operated knees, right knee
frog	Word2Vec	toad, bullfrog, amphibian, rana, turtle, salamander, caudiverbera, frogs, leptodactylid, pleurodema
	GloVe	rana, toad, amphibian, bullfrog, frogs, temporaria, laevis, xenopus, anuran, catesbeiana
	Ariadne	frogs, isolated frog, frog muscle, rana pipiens, anurans, hyla, anuran, tree frog, anuran species, hylid

Different corpus, different embedding

What is *young*?

WorldCat	people, children, adolescents, nobleman, christians, pianists, siblings, vietnamese, clergyman, housekeeper
Medline	adults, children, people, women, men, adulthood, infants, athletes, girls, leaves, patients, mania, boys, chicks, calves
Art library	people, children, persons, adults, lady, women, gentlemen, artists, readers, folks, americans, memorial, girls, architects
Astrophysics	stars, supernova, stellar, clusters, massive star clusters, brown dwarf

What about the temporal dimension?

- 20 million Medline articles published since 1977
- 1.5 million entities (subjects, authors, journals, words)
- 8 five-year periods
- Each subject is embedded in 8 chronological vector spaces
- Is there concept drift and can we detect it?

Most and least stable MeSH subjects

Most stable subjects	Least stable subjects
history 15th century	diagnostic techniques, surgical
history 18th century	chromium isotopes
history 17th century	shock, surgical
history 16th century	iodine isotopes
history 19th century	diagnostic techniques and procedures
thymoma	blood circulation time
history ancient	trauma nervous system
history medieval	cesium isotopes
rabies	liver extracts
history	macroglobulins

Subjects most related to “trauma nervous system”

1977-1982	anatomy regional, fracture fixation internal, bulgaria, piedra, surgery plastic, germany west, wound infection, carbuncle
1982-1987	l with hyperactivity, legionnaires disease, transfer p
1987-1992	l ction, orthopedic equipment, dermatomycoses, r
1992-1997	p proteins, emaciation, professional patient relation
1997-2002	defensive medicine, insurance liability, diagnostic dimethyl sulfate, medical errors, p protein
2002-2007	peripheral nervous system diseases, peripheral peripheral nerves, elbow, comorbidity, mother ch
2007-2012	peripheral nerve injuries, sciatic neuropathy, papilledema, sciatic nerve, peripheral nerves, nerve crush, neuroma, nerve regeneration, acute disease
2012-2017	mitochondrial dynamics, dental records, park7 protein human, persistent vegetative state, dnm1l protein human, platelet derived growth factor bb, dual specificity phosphatases, lingual nerve injuries, dental care

anatomy regional, fracture fixation internal, bulgaria, piedra, surgery plastic

defensive medicine, insurance liability, diagnostic errors, expert testimony, birth injuries,

Summary

- Semantic indexing helps to discover links between entities
- Links might have to be time stamped
- Free-text in metadata is a promising but challenging source
- No perfect algorithms yet but lots of on-going research