

From trees to graphs: Creating Linked Data from XML

Catherine Dolbear & Shaun McDonald
Content Architecture, Global Academic Business
Oxford University Press

Overview

-
- OUP and our business drivers
 - Approaches in the literature
 - Our publishing workflow and XML metadata
 - Modelling RDF graphs from XML trees
 - Semantic markup: RDFa and schema.org
 - Summary

Introduction to OUP

Meet the Press...

OXFORD
UNIVERSITY PRESS

OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

DATABASE

The Journal of Biological Databases and Curation

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Institution: OUP site access Sign in as Personal Subscriber

Oxford Journals > Life Sciences > Database > Volume 2013 > 10.1093/database/bat018

MalaCards: an integrated compendium for diseases and their annotation

Noa Rappaport¹, Noam Nativ¹, Gil Stelzer², Michal Twik¹, Yaron Guan-Colan², Tsippi Iny Stein¹, Iris Bahir¹, Frida Belinky¹, C. Paul Morrey¹, Marilyn Safran¹ and Doron Lancet¹

Author Affiliations

Corresponding author: Tel: +972-8-934-3188; Fax: +972-8-934-4108; Email: noa.rappaport@weizmann.ac.il

Received November 27, 2012.
Revision received February 7, 2013.
Accepted March 13, 2013.

Abstract

Comprehensive disease classification, integration and annotation are crucial for biomedical discovery. At present, disease compilation is incomplete, heterogeneous and often lacking systematic inquiry mechanisms. We introduce MalaCards, an integrated database of human

Home About What's new Contact us Subscriber services Help

OED

Oxford English Dictionary
The definitive record of the English language

ailurophobe, *n.*

View as: Outline | Full entry

Pronunciation: Brit. /aɪˈlʊərəʊ(ə)fəʊb/, /eɪˈl(j)ʊərə(ʊ)fəʊb/, /eɪljərə(ʊ) /aɪˈlʊərəˌfoʊb/, /eɪˈlʊərəˌfoʊb/

Forms: 19- aelurophobe, 19-aelurophobe, 19-ailurophobe, 19-ailur

Oxford Scholarly Editions Online

MAIN TEXT NOTES RESET PANELS

Main Text

UNIVERSITY PRESS SCHOLARSHIP ONLINE About What's New Partner Presses Subscriber Services Contact Us Take a Tour Help

Oxford Scholarship Online

Sign in not registered Sign up

Search

Search by My Subject Specializations: Beled...

My entries (1) | My searches (0)

Browse by Subject

- Biology
- Business and Management
- Classical Studies
- Economics and Finance
- History
- Law
- Linguistics
- Literature
- Mathematics
- Music
- Neuroscience
- Pathology
- Philosophy
- Physics
- Political Science
- Psychology
- Public Health and Epidemiology
- Religion
- Social Work
- Sociology

The Machinery of Criminal Justice

Stephanos Bibas

Abstract

Two centuries ago the criminal justice system was primarily run by laymen. In court, victims and defendants interacted face to face while lay jurors from the community sat in judgment. Jury trials passed moral judgment on crimes, vindicated victims and innocent defendants, denounced guilty defendants, and reconciled and healed wounded relationships. But over the last two centuries, lawyers have taken over the process, alienating victims and defendants and, in many cases, substituting a plea-bargaining system for voice of the jury. This lavishly illustrated machinery has garbaged efficiency, speed, process...

Keywords: criminal justice system, laymen, victims, defendants, lay jurors, jury trials, lay participation

Biographical information

Print ISBN-13: 9780195374482
Published to Oxford Scholarship Online: May 2012
DOI: 10.1093/acprof/oso/9780195374482.001.0001

Find in Library

Print in Wordfast

Google

Contents

Go to page: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

Front Matter

Chapter 1 The Long Drift from Morality Play to Machine

Chapter 2 *in* *opere*, Unresponsive Criminal Justice

OXFORD ISLAMIC STUDIES ONLINE ABOUT WHAT'S NEW LOGOUT

Oxford Dictionary of National Biography

Search Browse Themes Quick Search person Search Tips

People | Full text | References | Contributors | Images

Print Email Cite

Davidson, Emily Wilding

(1872-1913), suffragette, by Vera Di Campilii San Vito

References

Themes

Women's Social and Political Union

Other online resources

National Portrait Gallery
National Register of Archives

DNB archive


© Oxford University Press 2004-13

Home | About Oxford DNB | What's new | Subscriber services | Contact us | Help | Logout

QUR'ANIC STUDIES TIMELINES LEARNING RESOURCES

Previous Result Results Look It Up Highlight On / Off Next Result

Ablution




QUR'ANIC VERSE LOOKUP DATE CONVERTER

Subscriber Services | Contact Us | Help

QUR'ANIC STUDIES TIMELINES LEARNING RESOURCES

Previous Result Results Look It Up Highlight On / Off Next Result

Ablution



Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date;
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimmed,
And every fair from fair sometime declines,
By chance or nature's changing course untrimm'd:
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st;
Nor shall Death brag thou wand'rest in his shade,
When in eternal lines to time thou grow'st.
So long as men can breathe or eyes can see,
So long lives this, and this gives life to thee.

too hot nor too cold; of mild and equable temperature' (OED 3a)

- 4 lease temporary period of legal possession, limited by a date, or period of expiry. See 13.5 and 6 nn.
- 6 complexion (a) 'Countenance, face' (OED 4c); (b) 'Colour, visible aspect, look, appearance' (OED 5 transf.), as in *Richard II* 3.2.190-1: 'Men judge by the complexion of the sky | The state and inclination of the day'

Motivation and business drivers

- Search Engine Optimisation
 - Discoverability of our subscription content
 - “Index card” of XML metadata published open access
- Improvement of user journeys across multiple products
 - Dynamic links generated as search results
 - Static links e.g. *is Author Of*, has *Primary Topic* currently stored as XML documents

The screenshot shows the Oxford Index search results for the term 'chocolate'. The page layout includes a search bar at the top right, navigation links for 'Language', 'My Content (1)', and 'My Searches (0)', and a search icon. The main content area is titled 'OVERVIEW chocolate'. It features a 'QUICK REFERENCE' section with a paragraph of text: 'An alkaloid refined from cacao seeds, originally native to Central America and used by priests in pre-Columbian religious rituals; chocolate took Europeans by storm when brought back to Spain on Columbus's third voyage. Soon it was used everywhere in confectionary and drinks. It may be the most widely used confectionary in the world. It is not addictive like nicotine but some people ("chocoholics") experience periodic cravings. It induces feelings of comfort among many who consume it. Excessive consumption can contribute to obesity because of the high calorie content of most chocolate confectionary. It has no adverse health effects, and there is some evidence that dark chocolate helps to provide protection from coronary heart disease.' Below this is a 'From:' link to 'chocolate in A Dictionary of Public Health' and a 'Subjects:' link to 'medicine and health'. There is also a 'RELATED CONTENT IN OXFORD INDEX' section with a link to 'See all related items in Oxford Index'. At the bottom, there is a 'REFERENCE ENTRIES' section with a link to 'Chocolate in The Oxford Encyclopedia of Food and Drink in America'. On the right side, there is a sidebar with 'OXFORD UNIVERSITY PRESS - MAIN ACCIT' and 'RELATED OVERVIEWS' which lists 'cocoa', 'Valentine's Day', 'cookie', and 'brownie', along with a link to 'See all related overviews in Oxford Index'. At the bottom right, there is a section titled ''CHOCOLATE' CAN ALSO REFER TO...' which lists 'Alfredo 'Chocolate' Armenteros (b. 1928)', 'Blood & Chocolate (Elvis Costello album)', and 'Carolina Chocolate Drops'.

Approaches in the literature

What's been tried before

- MarkLogic
 - XQuery to construct triples from XML, linked using URIs
 - We follow this pattern using Digital Object Identifiers expressed as URIs
- BBC
 - Statistics and content in MarkLogic XML database
 - Journalists annotate assets according to an ontology, results stored in OWLIM triple store.
 - Content aggregated by combining SPARQL and XQuery e.g. "The league table for the English Premiership"
- Nature Publishing Group
 - Adobe XMP, a subset of RDF embedded in XML documents
 - Triple store enables integrated queries of all XML content distributed across the organisation



Oxford Index

Safari PubFactory platform



Product website



Product website



Metadata for products included on Oxford Index

Content + Product Metadata

PubFactory repository

Metadata Hub REST API

XML/Triple Store

Pre-ingestion layer

MarkLogic CMS

Metadata for products requested by Library Service

Metadata for all OUP Content

Link data

Link generation

Full Text

Metadata



Library Services, Aggregators



Product website

Content + product metadata

Metadata

Onix Data

Product Data

Metadata

High Wire



Product website

Metadata

CMS

Full Text



Product website

- Single vocabulary for metadata for all products
 - Originates from multiple sources with varying DTDs or none
 - MarkLogic, FileMaker, SQL server, even Excel spreadsheets
- Reuses some Dublin Core vocabulary, plus terms based on our own needs
- Links embedded in XML document or “stand-alone” OxMetaLinkML documents
 - Named predicates like “*is author of*”, “*is related to*”, “*is primary topic of*”
- Published as XML for externally-developed product website platform
 - Document-centric

Modelling RDF graphs

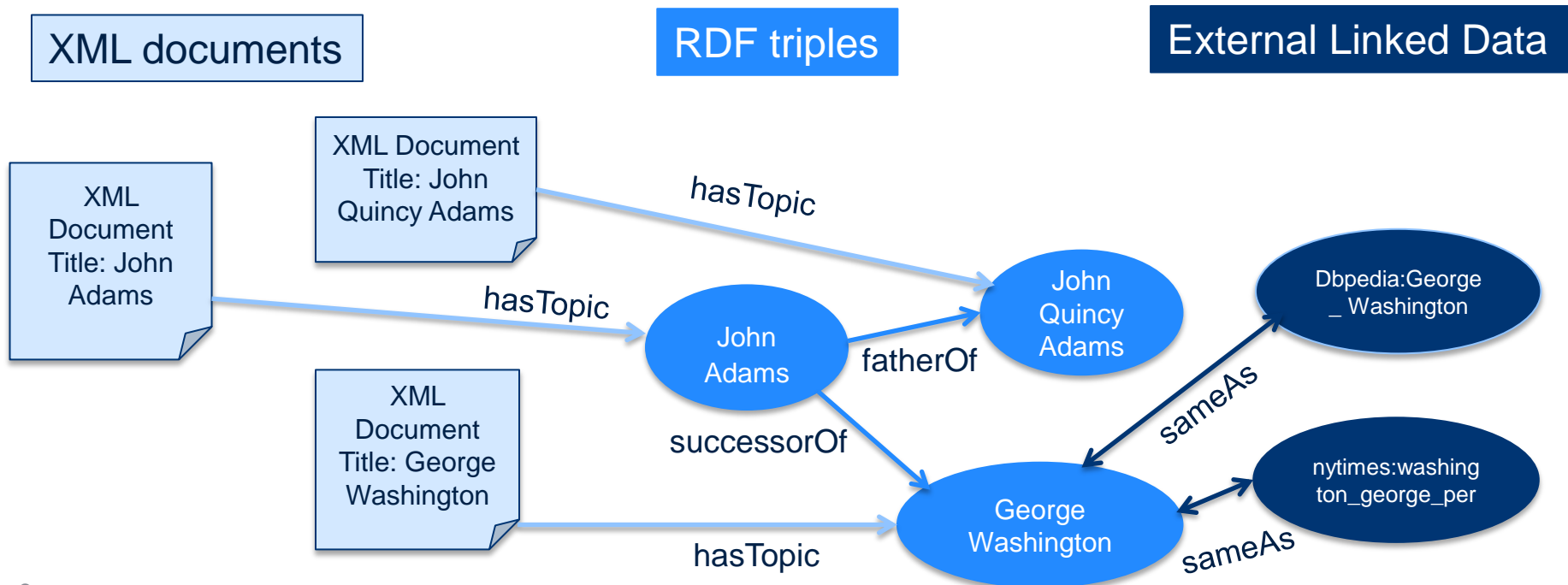
There is no order...

- XML: documents, elements, sequential order – *trees*
- RDF: relationships between concepts - *vertices and arcs*
 - Difficult to manipulate relationships in XML
- XML for content, RDF for metadata
- Our metadata includes abstracts and must be output to XML
- But as more concepts in the XML become linked in their own right and given identifiers, more can migrate to a graph model.

Bibliographic versus semantic metadata

Information versus meaning

- Bibliographic information (author, title, ISBN etc)
- Semantic or contextual information - what the document is about (academic subject, person, organisation etc)



RDF Data Model

- RDF is a data model (graph) not a syntax
- Use Turtle, not RDF/XML
 - Less verbose, less syntactic variation
 - Can concentrate on knowledge modelling
 - Element order and syntactic use of `rdf:Description` or `rdf:about` is irrelevant
- Better performance to generate inverse triples from SPARQL query rather than store explicitly or use inference

Examples

Turtle and SPARQL

DOI123 a oup:Document.

DOI123 foaf:hasTopic URI456.

URI456 oup:hasName "George Washington".

URI456 oup:hasSuccessor URI789.

URI789 oup:hasName "John Adams".

Examples

Turtle and SPARQL

DOI123 a oup:Document.

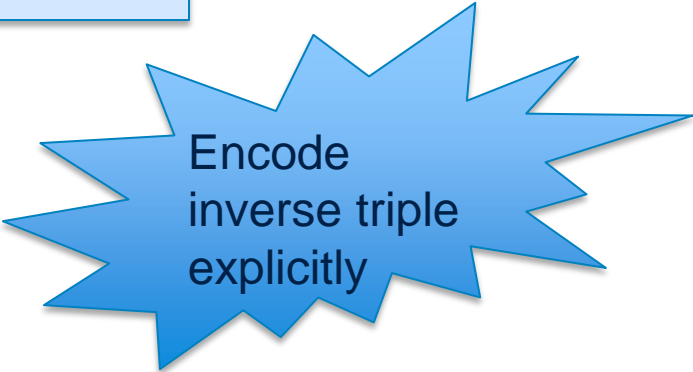
DOI123 foaf:hasTopic URI456.

URI456 oup:hasName "George Washington".

URI456 oup:hasSuccessor URI789.

URI789 oup:hasName "John Adams".

URI789 oup:isSuccessorOf URI456.



Encode
inverse triple
explicitly

Examples

Turtle and SPARQL

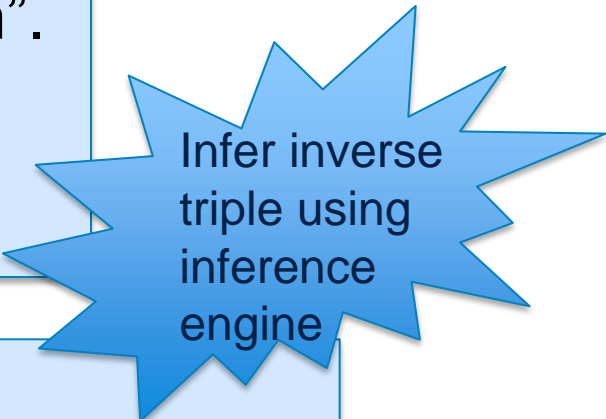
DOI123 a oup:Document.

DOI123 foaf:hasTopic URI456.

URI456 oup:hasName "George Washington".

URI456 oup:hasSuccessor URI789.

URI789 oup:hasName "John Adams".



Infer inverse
triple using
inference
engine

oup:hasSuccessor a rdf:Property.

oup:hasSuccessor owl:inverseOf oup:isSuccessorOf.

=> URI789 oup:isSuccessorOf URI456.

Examples

Turtle and SPARQL

```
DOI123 a oup:Document.
DOI123 foaf:hasTopic URI456.
URI456 oup:hasName "George Washington".
URI456 oup:hasSuccessor URI789.
URI789 oup:hasName "John Adams".
```

```
CONSTRUCT {?subject oup:isSuccessorOf URI456}
WHERE {
    URI456 oup:hasSuccessor ?subject.
}
```

Generate inverse
triple as query
result

Result:
URI789 oup:isSuccessorOf URI456.

Reification

Information about the triples

- Accuracy of the link, date of creation, approval status etc.
- Can store a fourth piece of information in RDF by:
 - Named graphs aka “quads”. More suited to groups of triples
 - Assign a URI to each triple and treat as a resource using RDF *reification* vocabulary

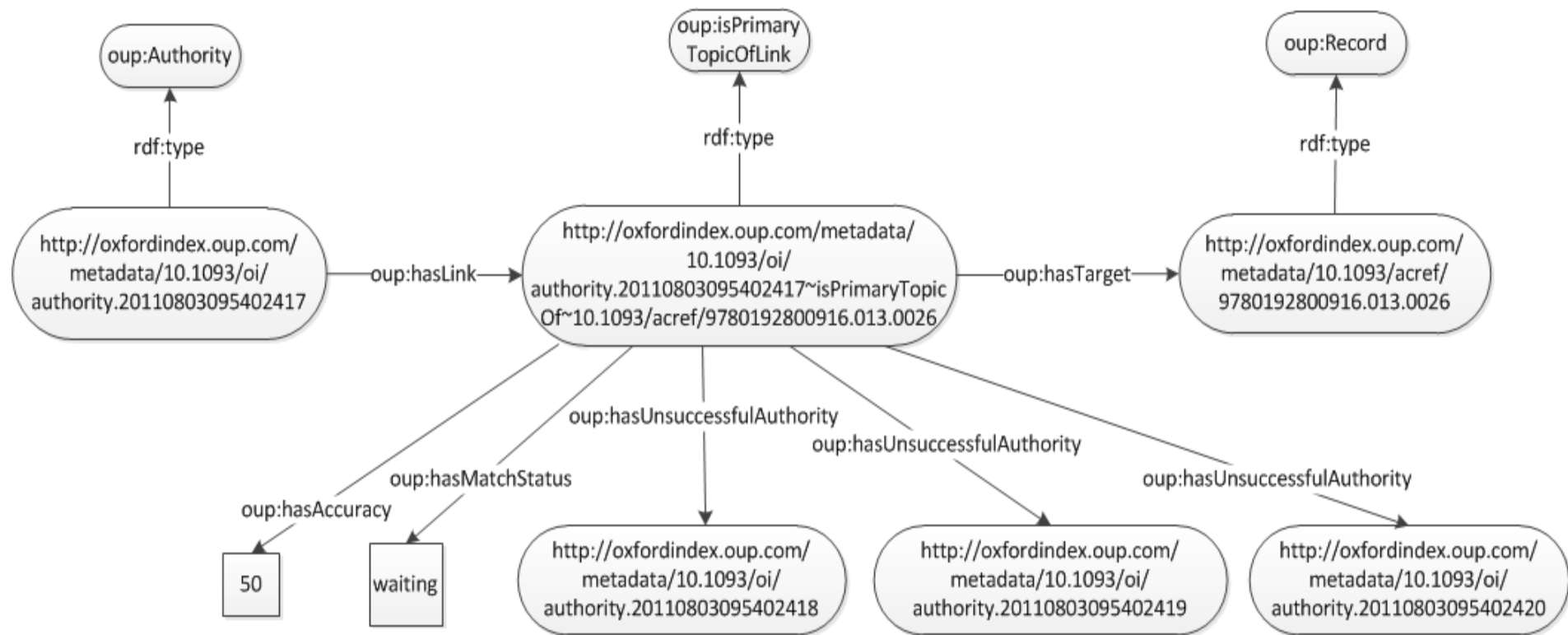
```
<URI20110803100243337> oup:hasOccupation “President of the United States”.
```

```
<Statement12345> a rdf:Statement;  
    rdf:subject <URI20110803100243337>;  
    rdf:predicate oup:hasOccupation;  
    rdf:object “President of the United States”.
```

```
<Statement12345> oup:isValidFrom “20 January 2009”.
```

Reification using RDFS Classes

Simpler queries; better performance



Linked Data

principles for connecting information on the web

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 related things

- Connections across content, not just documents
- Distinguishes between a *document about* Barack Obama, and the man himself
- At the moment, our DOIs provide documents, not data

Business cases for Linked Data

Where's the money?

- Internal benefits for using RDF:
 - Storing links between XML documents
 - Using external RDF data to augment our metadata (e.g. OBO ontology to identify gene names in abstracts)
- ROI from publishing OUP metadata as Linked Data less clear
- Could be used to supply metadata to library services and aggregators (e.g. EBSCO, Summon)
- Business models: branding, freemium, **traffic model**
 - First step to publish RDF as embedded markup

RDFa and schema.org markup

Embedding RDF in HTML

- Improves click-through rate (30% reported by BestBuy) as search results more eye-catching

```
<div vocab="http://schema.org/"
  typeof="Person"
  about="http://oxfordindex.oup.com/
view/10.1093/oi/authority.20110803100243337">
  <span property="name">Barack Obama</span>
</p>
  <span property="jobTitle">American Democratic
statesman</span> </p>
  born <span property="birthDate">4 August
1961</span> </p>
</div>
```

Barack Obama



plus.google.com

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. [Wikipedia](#)

Born: August 4, 1961 (age 50), [Honolulu](#)

Full name: Barack Hussein Obama II

Net worth: US\$ 10.5 million (2010)
[celebritynetworth.com](#)

Education: [Harvard Law School](#) (1988–1991),
[Columbia University](#) (1983), [More](#)

Children: [Natasha Obama](#), [Malia Ann Obama](#)

Books: [Dreams from My Father](#), [The Audacity of Hope](#), [Of Thee I Sing](#), [More](#)

RDFa versus schema.org

- RDFa allows for richer descriptions
 - Can provide our full metadata “under the hood”
- But schema.org fully supported by major search engines
 - We could use CreativeWork schema (Book, Article concepts) as well as Person
- Drawback is that only simple markup can be used
 - Can introduce semantic mismatch – is “American democratic statesman” really a job title?
 - Not a full alternative to an API or Linked Data publication

Summary

Our journey from XML to Linked Data

- We're still in the early days...
- Internal business case for semantic technologies and link generation (SEO, user journeys) is much stronger than for Linked Data publication itself
- XML for documents, RDF for relationships
 - How much of our metadata should we store as RDF?
- Is our experimental architecture of an XML store for documents and a triple store for links the most performant?