THE ROLE OF METADATA IN THE SECOND MACHINE AGE

DC-2016 / KØBENHAVN / 13 OCTOBER

Bradley P. Allen
Chief Architect, Elsevier
@bradleypallen
DUBLIN CORE IN ITS THIRD DECADE

A basic description mechanism for digital information that:

- can be used in all domains
- can be used for any type of resource
- is simple, yet powerful
- can be extended and can work with specific solutions.

Since its early days in the mid-1990s, DCMI's founding principle has been the discovery and management of resources through metadata across the boundaries of information silos on the Web and within intranets.

THE IMPACT OF DC AND DCMI

• Arguably the most impactful of many efforts to bring to bear the library science approach to information organization on the Web at large

• Simple, easy to understand, easy to apply, well documented

• While the user experience of discovery has come to be dominated by search engines such as Google, metadata standards are pervasive in the infrastructure of content curation and management, and underpins search infrastructure

• All of this has been a huge contribution to the Web

• It established the approach to bringing the Semantic Web and its associated vocabularies and ontologies online as the Linked Open Data Cloud
THE NEXT 20 YEARS: THE SECOND MACHINE AGE

The second machine age will be characterized by countless instances of machine intelligence and billions of interconnected brains working together to better understand and improve our world.

MACHINE INTELLIGENCE IS ON THE RISE
WHAT CAN MACHINE INTELLIGENCE DO TODAY?

If there’s a task that a normal person can do with less than one second of thinking, there’s a very good chance we can automate it with deep learning.

Andrew Ng, Chief Scientist, Baidu (lecture at Bay Area Deep Learning School, Stanford, CA, September 24, 2016)
# MACHINE INTELLIGENCE IN USE

<table>
<thead>
<tr>
<th>Application areas</th>
<th>Specific examples</th>
<th>Recent M&amp;A and hiring</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Speech understanding</td>
<td>• &quot;Hummingbird&quot; release contextual search</td>
<td>• Deepmind</td>
</tr>
<tr>
<td>• Web search</td>
<td>• Google Now intelligent assistant</td>
<td>• Dark Blue Labs</td>
</tr>
<tr>
<td>• Image search</td>
<td>• Knowledge GraphVault</td>
<td>• Vision Factory</td>
</tr>
<tr>
<td>• Machine translation</td>
<td>• Spam filtering in Gmail</td>
<td>• Timeful</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Self-driving cars</td>
<td>• DNNResearch</td>
</tr>
<tr>
<td>• Logistics</td>
<td></td>
<td>• Hinton (U. Toronto)</td>
</tr>
<tr>
<td>• Speech understanding</td>
<td>• Bing contextual search</td>
<td>• Merging Bing, Cortana and MSR into ~5,000-person AI division</td>
</tr>
<tr>
<td>• Cloud services</td>
<td>• Cortana intelligent assistant</td>
<td></td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Azure Machine Learning Services</td>
<td></td>
</tr>
<tr>
<td>• Speech understanding</td>
<td>• Satori Knowledge Base</td>
<td></td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Microsoft Cognitive Services</td>
<td></td>
</tr>
<tr>
<td>• Cloud services</td>
<td>• Turi</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• iOS9 Proactive Suggestions</td>
<td></td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Siri intelligent assistant</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Cloud services</td>
<td>• iOS9 Proactive Suggestions</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Healthcare decision support</td>
<td>• Watson Discovery for Life Sciences</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Watson Discovery Services</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Cloud services</td>
<td>• Watson Discovery Services</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Healthcare decision support</td>
<td>• Watson Health Cloud</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Watson for Health</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face detection &amp; recognition</td>
<td>• Watson Discovery Services</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Facebook Open Graph</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Facebook Graph Search</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Establishing AI division in Seattle</td>
</tr>
<tr>
<td>• Face detection &amp; recognition</td>
<td>• Facebook Open Graph</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face detection &amp; recognition</td>
<td>• Facebook Open Graph</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face detection &amp; recognition</td>
<td>• Facebook Open Graph</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face detection &amp; recognition</td>
<td>• Facebook Open Graph</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Face recognition in Facebook, Instagram</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Face detection &amp; recognition</td>
<td>• Facebook Open Graph</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Personalization and contextual search</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
<tr>
<td>• Question answering</td>
<td>• Facebook Graph Search</td>
<td>• Turi</td>
</tr>
</tbody>
</table>
WHAT DOES THIS MEAN FOR THE FUTURE OF METADATA?

• Our focus as a community has been on helping people find and use information on the Web
• Metadata standards and machines have been a means to that end
• The emergence of machine intelligence and machine reading in the second machine age will make it even easier to automate the production of metadata to help people find, filter and organize information
• Things information workers do in less than a second that we can train machines to do: read a page for concepts and facts, recognize the type of image on a page, …
My work is moving towards a new field; what should I know?

• Journal articles, reference works, profiles of researchers, funders & institutions
• Recommendations of people to connect with, reading lists, topic pages

How should I treat my patient given her condition & history?

• Journal articles, reference works, medical guidelines, electronic health records
• Treatment plan with alternatives personalized for the patient

How can I master the subject matter of the course I am taking?

• Course syllabus, reference works, course objectives, student history
• Quiz plan based on the student’s history and course objectives
THE GROWTH OF SCIENCE COMPlicATES OUR EFFORTS
ANSWERS ARE ABOUT THINGS, NOT JUST WORKS

Why shouldn’t a search on an author return information about the author, including the author’s works? Where was the author born, when did she live, what is she known for? … All of this is possible, but only if we can make some fundamental changes in our approach to bibliographic description. … The challenge for us lies in transforming what we can of our data into interrelated “things” without overindulging that metaphor.

KNOWLEDGE GRAPHS AND MACHINE READING TURN CONTENT INTO ANSWERS

• Knowledge graphs are "graph structured knowledge bases (KBs) which store factual information in form of relationships between entities" (Nickel, M., Murphy, K., Tresp, V. and Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. arXiv:1503.00759v3)

• Knowledge graphs are metadata evolved beyond the focus on the work, linking people, concepts, things and events

• Knowledge graphs organize data extracted from content through machine reading so that queries can provide answers
ELSEVIER: KNOWLEDGE GRAPHS FOR LIFE SCIENCES

Biological Pathways extracted via semantic text mining

Bioactivities through text analysis

Chemical Structures And Properties

A upregulates B
B upregulates C
C increases disease D
A ⇔ B ⇔ C ⇔ D

\[ \text{IC}_{50} 6.3\text{nM}, \text{kinase binding assay} \]

10mM concentration

Bioactivities

IC\(_{50}\) 6.3nM, kinase binding assay
10mM concentration

Chemical Structures

And Properties

Normalizing vocabularies required: proteins, diseases, drugs, chemicals

EMTREE

NCBI, Uniprot

InChi, Name

ReaxysTree, Structures
ELSEVIER: KNOWLEDGE GRAPHS FOR LIFE SCIENCES
ELSEVIER: KNOWLEDGE GRAPHS FOR HEALTHCARE
{  
  "@context": [
    "http://vtw.elsevier.com/metadata/context.jsonld",
    {"eaf": "http://vtw.elsevier.com/data/ns/Formats/Annotation",
     "oca": "http://ontochem.com/ns/ann"}
  ],
  "@type": "eaf:AnnotationDocument",
  "eaf:annotatedSourceInfo": {
    "dct:identifier": "E-602148748",
    "dct:title": "A rare mutation in MYH7 gene occurs with overlapping phenotype",
    "dct:type": "http://data.elsevier.com/vocabulary/ElsevierContentTypes/1.3"
  },
  "eaf:annotationProvenance": {
    "dct:format": "application/xhtml+xml",
    "cef:enrichmentTool": "OCMiner",
    "cef:enrichmentToolVersion": "RxT edition 2.2.Milestone6",
    "cef:dictionary": "ReaxysTree",
    "cef:dictionaryVersion": "13.0",
    "prov:generatedAtTime": "2016-02-16T16:11:07Z"
  }
},
"eaf:annotations": [
  {
    "eaf:localID": "1",
    "eaf:sourceText": "beta-myosin",
    "eaf:conceptID": "",
    "eaf:annotationType": "concept",
    "eaf:annotationScope": "text range",
    "eaf:range": "2726-2737",
    "eaf:confidence": "77",
    "eaf:relevance": "72",
    "oca:d": "chemCompound",
    "oca:CNType": "Compound",
    "oca:CNHasStructure": "NO",
    "oca:CNSubType": "Trivial"
  },
  {
    "eaf:localID": "2",
    "eaf:sourceText": "liquid",
    "eaf:conceptID": "817100004230",
    "eaf:annotationType": "concept",
    "eaf:annotationScope": "text range",
    "eaf:range": "7448-7454",
    "eaf:confidence": "65",
    "eaf:relevance": "22",
    "oca:d": "reaxystree",
    "oca:pn": "liquid"
  },
  ...
}
ELSEVIER: MACHINE LEARNING FOR LINK DISCOVERY

Diseases 2791370: Glaucoma

- Contact dermatitis
- Primary open-angle glaucoma
- Oxidative stress
- Retinal ganglion cells

Glaucoma has been documented to cause contact dermatitis and is assessed through evaluation of primary open-angle glaucoma. Treatment involves oxidative stress to the retinal ganglion cells.

Content

- Open Information Extraction
- Surface form relations
- Universal schema
- Factorization model
- Predicted relations

Taxonomy

- 14M articles from Science Direct
- 475M facts
- 3.3M facts

Knowledge graph

- 49M facts
- 920K concepts from EMMeT

Matrix Construction

- r = 176
- p = 83
- 83 x 176 sparse binary-valued matrix with 366 entries
- 83 x 176 real-valued matrix with 14,608 entries

Entity Resolution

- 381593 diseases
- 5415395 procedures
- 5216597 procedures
- 8164986 diseases
- 8002098 anatomy

Curation

- Glaucoma developed many years after chronic inflammation of uveal tract is considered the second leading cause of functional visual field loss.
- Chronic inflammation of uveal tract can appear soon in family history of glaucoma, age over 40, chronic inflammation of uveal tract.
- Functional visual field loss.
- Primary open-angle glaucoma.
- Functional visual field loss.
- Functional visual field loss.

Factorization

- Contributed to functional visual field loss.
- Contributed to functional visual field loss.
- Contributed to functional visual field loss.

Predicted relations

- Functional visual field loss.
- Functional visual field loss.
- Functional visual field loss.
- Functional visual field loss.
ELSEVIER: DEEP LEARNING FOR IMAGE ANNOTATION
MOVING FORWARD: QUESTION ANSWERING AS AN AI-COMPLETE PROBLEM

Question answering (QA) is a complex natural language processing task which requires an understanding of the meaning of a text and the ability to reason over relevant facts. Most, if not all, tasks in natural language processing can be cast as a question answering problem …

I really believe that the key battleground in any industry is that of its knowledge graph. Google has it for media/advertising, Netflix has it for filmed entertainment, Uber has it for inner city transportation, Facebook has it across social media as well as messaging and the multiples speak for themselves.

Tony Askew, Founder/Partner at REV (personal communication, September 29, 2016)
METADATA STANDARDS ARE AND WILL CONTINUE TO BE FOUNDATIONAL FOR KNOWLEDGE GRAPHS

• Semantic Web and Linked Data standards are the consensus approach for representing and sharing knowledge graphs over the Web

• There is a single thread from the establishment of the Dublin Core through Open Linked Data to the emergence of knowledge graphs
  • Linked data principles
  • Standard vocabularies, taxonomies and ontologies
BUT WHAT DO METADATA STANDARDS DO FOR MACHINE READING?

- Machines’ proficiency in constructing knowledge graphs from text, audio, images and video will depend on our ability to train them effectively to read information from the Web

- How machines read the Web today
  - Crawling and indexing Web resources, possibly semantically tagged (e.g. using schema.org)
  - Find-and-follow crawling of open linked data resources for ontology and data sharing and reuse
  - Programmatic access to APIs mediated through HTTP/S and other Internet protocols
THE SEMANTIC WEB WAS INTENDED FOR MACHINE READING

... that's the real idea behind the Semantic Web: letting software use the vast collective genius embedded in its published pages.

BUT THE SEMANTIC WEB IS BUILT FOR PEOPLE, NOT MACHINES

- The Semantic Web is largely a logicist take on the way knowledge is to be represented.
- The latest advances in machine intelligence are based on a connectionist approach to knowledge representation.
- There is a gap between how knowledge is represented in the Semantic Web and what deep learning is exploiting to such good effect.
- The Semantic Web is silent about how machines can become better readers, and hence better partners in the second machine age.
- How will we evolve metadata standards to better accommodate machines?
MACHINE READING IS ENABLED BY MACHINE LEARNING

Programming

Machine learning

CPU

GPU

algorithm

learning architecture

data

input

model

output
MACHINES SEE THINGS DIFFERENTLY THAN PEOPLE

Figure 1: The hex dump represented on the left has more information contents than the image on the right. Only one of them can be processed by the human brain in time to save their lives. Computational convenience matters. Not just entropy.

MACHINES LEARN THINGS DIFFERENTLY THAN PEOPLE
MACHINE LEARNING DATASETS AND MODELS ARE BECOMING PART OF THE WEB

• Machines need lots and lots of data to learn how to read

• Datasets with ad-hoc formats are being made openly available
  • Open Images “~9 million URLs to images that have been annotated with labels spanning over 6000 categories” (The Open Images Dataset. (n.d.). Retrieved September 29, 2016, from https://github.com/openimages/dataset.)

• Standard architectures for machine (deep) learning are being released as open source
  • Dense neural networks for classification
  • Convolutional neural networks for image, audio and video recognition
  • Recurrent neural networks for sequence processing and generation

• Advances in the field are being published quickly and transferred to industrial application just as quickly
VOCABULARIES ARE SETS OF VECTOR EMBEDDINGS

Figure 3: Emoji vector embeddings, projected down into a 2-dimensional space using the t-SNE technique. Note the clusters of similar emoji like flags (bottom), family emoji (top left), zodiac symbols (top left), animals (left), smileys (middle), etc.

TRAINING DATASETS ARE GROWING IN VOLUME AND COVERAGE

Figure 2: The progression of datasets for image and video understanding tasks. Large datasets have played a key role for advances in both areas.

MODELS ARE BECOMING REUSABLE DATA RESOURCES THEMSELVES
THE ROLE OF METADATA IN THE SECOND MACHINE AGE – DC-2016 / KØBENHAVN / 13 OCTOBER

THE CHALLENGE FOR THE METADATA COMMUNITY: LINKED DATA THAT MACHINES CAN LEARN FROM

• We should be able to keep the core linked data principles that allow us to leverage Web architecture
  • URI as identifiers support (re)use of data in place (e.g. as in Google Open Images)
• We need to understand whether linked data formats need to change to support the needs of machine readers
  • The need for n-ary relations
  • The need for efficient indexing
• We need to investigate how vocabulary management can adapt to support the needs of machine reading
  • Vocabularies define what machines can recognize
  • Vocabularies need to accommodate lexical and vector representations together
WORK AT ELSEVIER: RESEARCH DATA MANAGEMENT
As machines become increasingly capable of general-purpose language understanding, the burden of effort in building machine intelligences will shift from software engineering to the acquisition, organization and curation of training content and data.
SAVE THE TIME OF THE MACHINE READER

Perhaps this law is not so self-evident as the others. Nonetheless, it has been responsible for many reforms in library administration and has a great potentiality for effecting many more reforms in the future.

SCHOLARLY PUBLISHING IN THE SECOND MACHINE AGE

Publishing content for people

Publishing data for machines

context

content

learning architecture

data

GPU

model

knowledge
IN SUMMARY: THE ROLE OF METADATA IN THE SECOND MACHINE AGE

• We need to save the time of the (human) reader by providing answers, not content

• This requires a shift from document-centric to entity-centric management of knowledge

• Knowledge graphs are "vehicles of communication" between people and machines and the battleground on which dominance in markets will be established

• We can use machine reading to build knowledge graphs

• The Semantic Web falls short in helping machines learn how to read

• Metadata standards to support machine (deep) learning should receive attention from the metadata community

• Our goal: making information resources on the Web discoverable and comprehensible for both people and machines
THANK YOU

Bradley P. Allen
Chief Architect, Elsevier
@bradleypallen
IMAGE ATTRIBUTIONS

- p. 2: https://twitter.com/stuartweibel
- p. 3: https://vimeo.com/66143280
- p. 5: http://www.prospectmagazine.co.uk/other/erik-brynjolfsson-andrew-mcafee
- p. 6: http://media.wix.com/ugd/142eb4_7581cfcf099e4e31a52599315f77c648.pdf
- p. 6: http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
- p. 7: https://www.wired.com/2013/05/neuro-artificial-intelligence/
- p. 11: https://twitter.com/karencoyles/status/784810709641158656
- p. 23: https://twitter.com/tonyaskew
- p. 38: https://plus.google.com/photos/112452283652920158570/album/6059593062673732337/6059593180461846098

Save corporate logos on p. 8, all other images © 2016 Elsevier, Inc.