Materials Semantic Web

An Open Community of people and machines sharing knowledge

Peter Murray-Rust
University of Cambridge
Open Knowledge Foundation

Louisiana State University, US, 2013-06-07

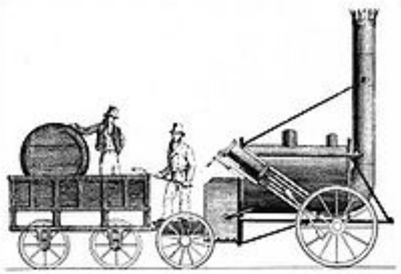
Unilever Cambridge
Centre for Molecular Informatics

Open Knowledge Foundation

PANTON ARMS
PANTON ARMS
GREEN

Overview

- Semantics will give us MUCH better, faster, more usable, scientific knowledge
- The revolution is comparable to railways, telephones.
- We scientists must work together to create the tools…and network
- PLANNING INVESTMENT



Themes

- WE can will change the world
- Formalization of knowledge is a community activity
- Let’s consider some communal problems in the next 2 hours
- Open Science means better science
- We want OUR suggestions

Hackathons

PMR Timeline for Semantic Materials

- 1994 1st WWW Conference, Chemical MIME
- 1994 Chemical Markup Language (HenryRzepa, PMR)
- 2001 UK eScience programme, eMinerals
- 2005 Materials Grid (Martin Dove group)
- 2006 Blue Obelisk (Open Source chemistry)
- 2006 Polymer Informatics (Unilever, Nico Adams)
- 2009 Chem4Word, OREChem (Microsoft Research)
- 2011 PNNL meetings and visit
- 2012 Semantic Physical Science (Cambridge)
- 2013 CSIRO meetings and visit
- 2013 LSU LA-Sigma

The Semantic Web

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

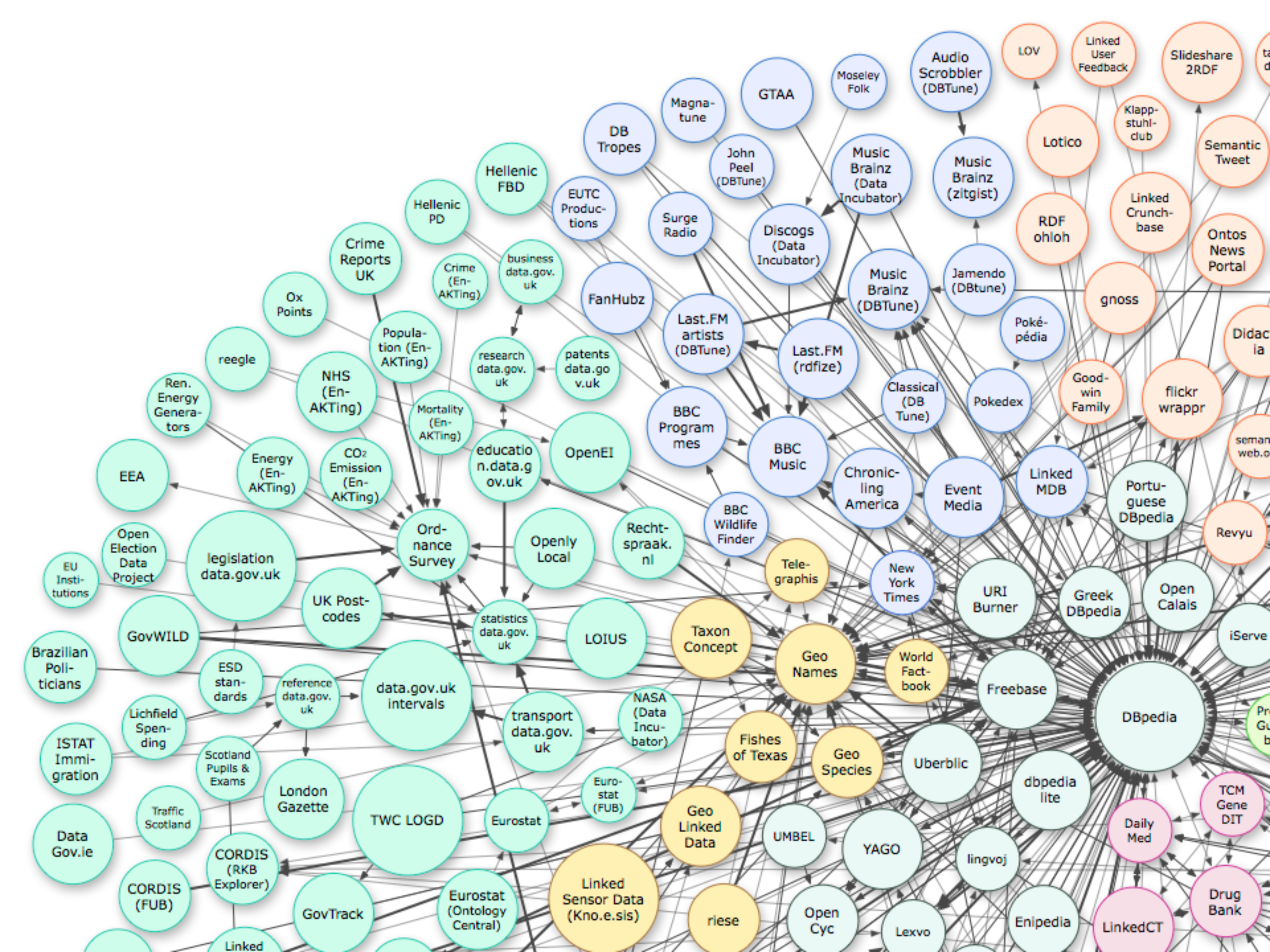
Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001

Semantic web (2006)

- I think maybe when you've got an overlay of **scalable vector graphics** [...] on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an **unbelievable data resource.**
  Tim Berners-Lee, A 'more revolutionary' Web (2006)

Linked Open Data – the world's knowledge

- http://upload.wikimedia.org/wikipedia/commons/3/34/LOD_Cloud_Diagram_as_of_September_2011.png


LOV
Linked User Feedback
Slideshare 2RDF
Audio Scrobbler (DBTune)
Moseley Folk
GTAA
Magna-tune
DB Tropes
John Peel (DBTune)
Music Brainz (Data Incubator)
Music Brainz (zitgist)
Lotico
Klapp-stuhl-club
Semantic Tweet
Hellenic FBD
Hellenic PD
EUTC Produc-tions
Surge Radio
Discogs (Data Incubator)
RDF ohloh
Linked Crunch-base
Ontos News Portal
Crime Reports UK
Crime (En-AKTing)
business data.gov.uk
FanHubz
Last.FM artists (DBTune)
Music Brainz (DBTune)
Jamendo (DBtune)
gnoss
Didac ia
Ox Points
Popula-tion (En-AKTing)
research data.gov.uk
patents data.gov.uk
Last.FM (rdfize)
Poké-pédia
reegle
NHS (En-AKTing)
Mortality (En-AKTing)
Classical (DB Tune)
Pokedex
Good-win Family
flickr wrappr
Ren. Energy Genera-tors
BBC Program mes
BBC Music
OpenEI
educatio n.data.g ov.uk
Energy (En-AKTing)
CO2 Emission (En-AKTing)
EEA
Chronic-ling America
Event Media
Linked MDB
Portu-guese DBpedia
seman web.o
BBC Wildlife Finder
Recht-spraak. nl
Revyu
Ord-nance Survey
Openly Local
Open Election Data Project
EU Insti-tutions
legislation data.gov.uk
Tele-graphis
New York Times
URI Burner
Greek DBpedia
Open Calais
UK Post-codes
statistics data.gov. uk
LOIUS
Taxon Concept
Geo Names
GovWILD
iServe
World Fact-book
Freebase
Brazilian Poli-ticians
ESD stan-dards
reference data.gov. uk
data.gov.uk intervals
DBpedia
NASA (Data Incu-bator)
transport data.gov. uk
Lichfield Spen-ding
Pr Gu b
Fishes of Texas
Geo Species
Uberblic
ISTAT Immi-gration
Scotland Pupils & Exams
dbpedia lite
TCM Gene DIT
London Gazette
Euro-stat (FUB)
Traffic Scotland
TWC LOGD
Eurostat
Geo Linked Data
Daily Med
UMBEL
YAGO
lingvoj
Data Gov.ie
CORDIS (RKB Explorer)
CORDIS (FUB)
Linked Sensor Data (Kno.e.sis)
GovTrack
Eurostat (Ontology Central)
riese
Open Cyc
Lexvo
Enipedia
LinkedCT
Drug Bank
Linked

The scientist's amanuensis

- *"The bane of my life is doing things I know computers could do for me" (Dan Connolly, W3C)*

Example: A semantic amanuensis could

- Give me a daily digest of zeolite papers
- Extract all the structures from them
- Submit those structures to GULP and NWChem
- Compare the results statistically
- Preserve and distribute the complete operation
- Prepare the results for publication

The semantic web is having a personal amanuensis

I'm AMI. I can't think, but I am very good at doing what I'm told

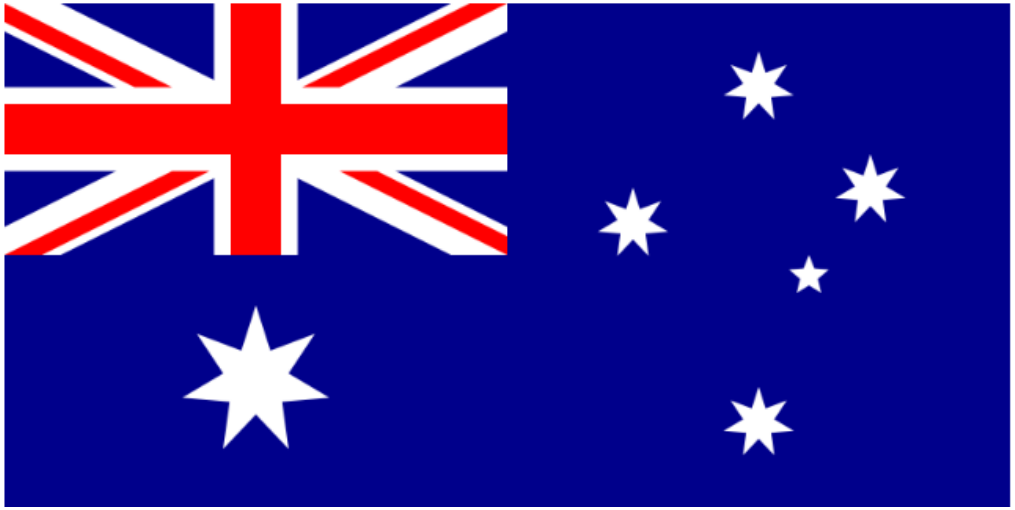
Connolly Challenge

- *The bane of my life is doing things I know computers could do for me" (Dan Connolly, W3C)*

NOW! Identify a single task where you think computers could save you significant time.

(This is NOT related to raw cpu power, but new software and information).

Scalable Vector Graphics (SVG)



Human-friendly

Automatic!

Machine-friendly

```
<?xml version="1.0" encoding="UTF-8"?>
<svg xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink" width="1280" height="640" viewBox="0 0
30240 15120">
<defs id="defs6">
<polygon points="0,-9 1.735535,-3.6038755 7.0364833,-5.6114082 3.8997116,-0.89008374 8.7743512,2.0026884 3.1273259,2.4939592
3.9049537,8.1087198 0,4 -3.9049537,8.1087198 -3.1273259,2.4939592 -8.7743512,2.0026884 -3.8997116,-0.89008374
-7.0364833,-5.6114082 -1.735535,-3.6038755 0,-9 " id="Star7"/>
</defs>
<path d="M 0,0 L 30240,0 L 30240,15120 L 0,15120 L 0,0 z" style="fill:#00008b"/>
<use transform="matrix(252,0,0,252,7560,11340)" id="Commonwealth_Star" style="fill:#fff" xlink:href="#Star7"/>
<use transform="matrix(120,0,0,120,22680,12600)" id="Star_Alpha_Crucis" style="fill:#fff" xlink:href="#Star7"/>
<!– snipped →
217,2520 L 10080,2520 L 15120,0 z" id="Red_Diagonals" style="fill:red"/>
<use transform="matrix(-1,0,0,-1,15120,7560)" id="Red_Diagonals_Rotated" style="fill:red" xlink:href="#Red_Diagonals"/>
</svg>
```

Mathematics Markup Language

MathML

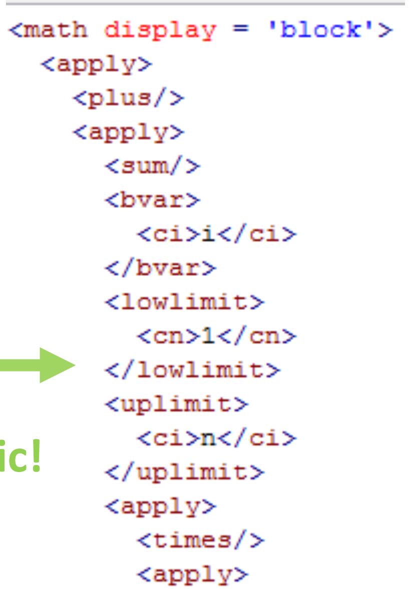
Energy of c.c.p lattice of argon

$$\sum_{i=1}^{n}\left(\mathbf{b}_i \times (\mathbf{a}_i)^{-\frac{s}{2}}\right) + \frac{pi\times\sqrt{32}}{(s-3)\times\left(\frac{3}{\sqrt{32\times pi}}\left(1+\sum_{i=1}^{n}\mathbf{b}_i\right)\right)^{\frac{s-3}{3}}}$$

Automatic!

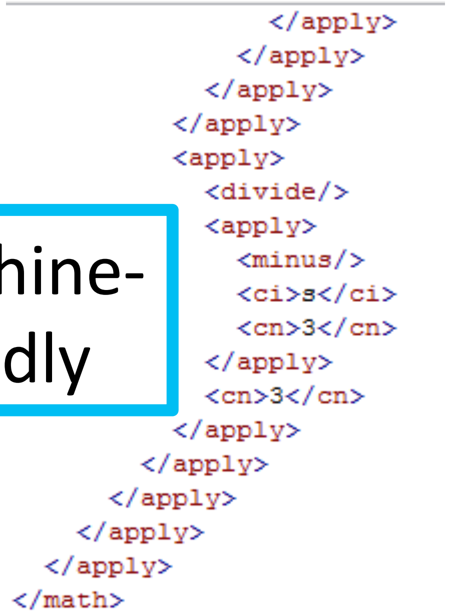
Human-friendly

Many editors and tools exist
We used MathWeaver



```
<math display = 'block'>
  <apply>
    <plus/>
    <apply>
      <sum/>
      <bvar>
        <ci>i</ci>
      </bvar>
      <lowlimit>
        <cn>1</cn>
      </lowlimit>
      <uplimit>
        <ci>n</ci>
      </uplimit>
      <apply>
        <times/>
        <apply>
```

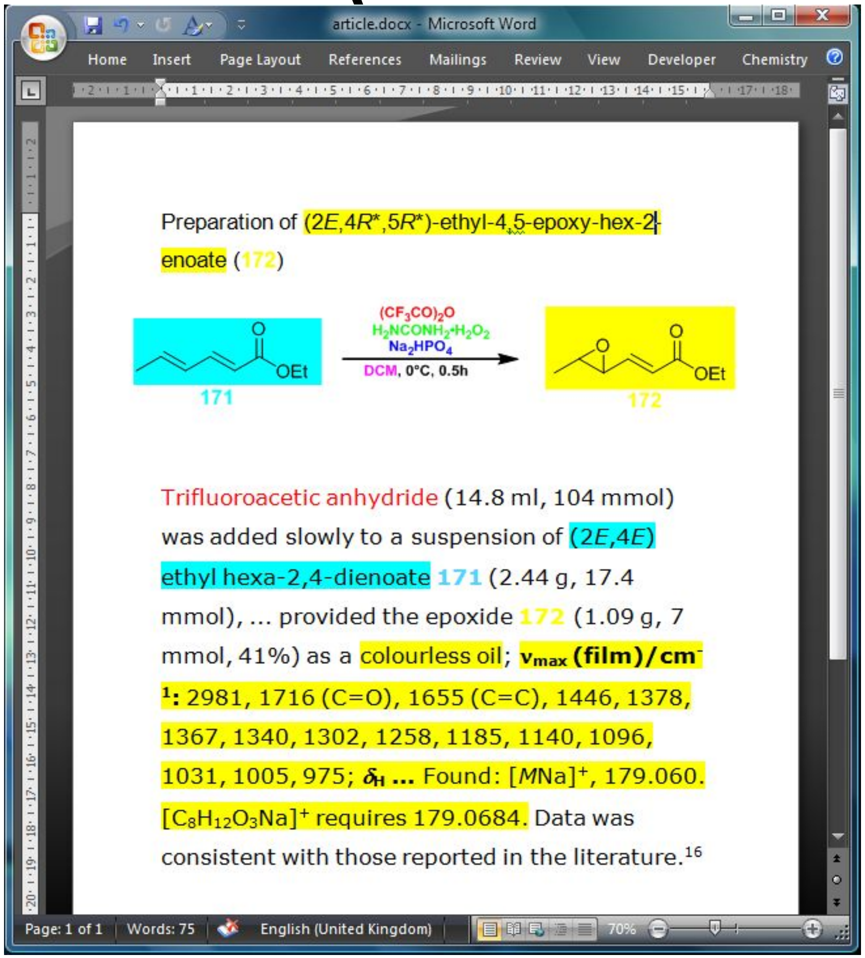
4 pages clipped



```
                </apply>
              </apply>
            </apply>
          </apply>
          <apply>
            <divide/>
            <apply>
              <minus/>
              <ci>s</ci>
              <cn>3</cn>
            </apply>
            <cn>3</cn>
          </apply>
        </apply>
      </apply>
    </apply>
  </apply>
</math>
```

Machine-friendly

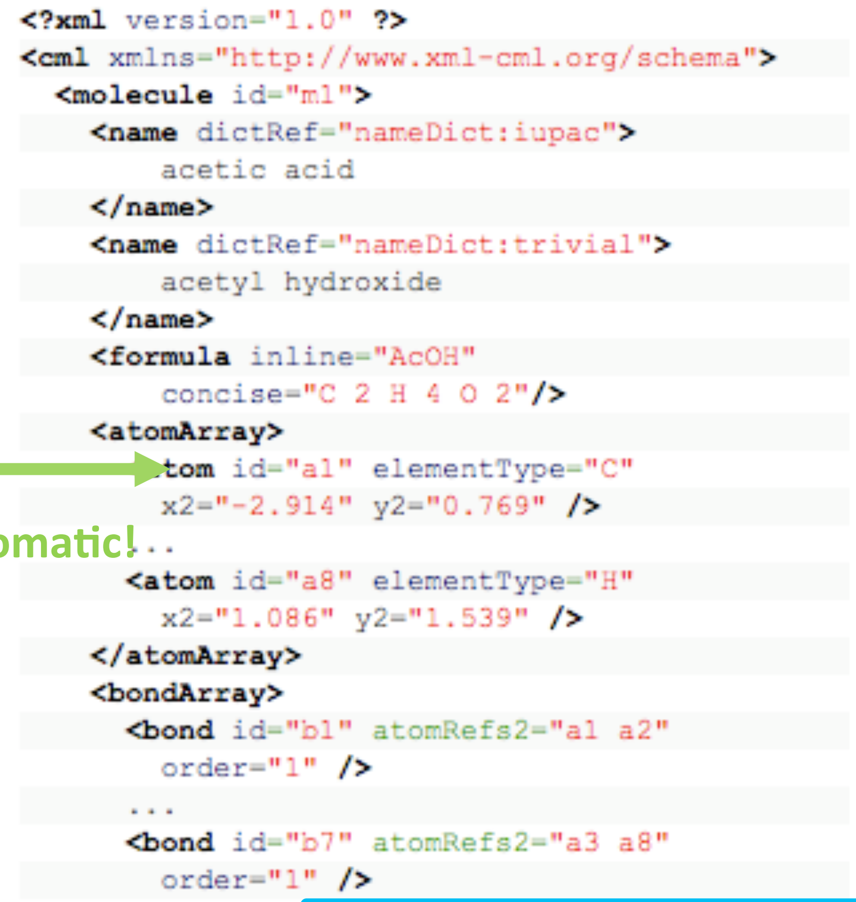
CML (Chemical Markup Language)



Preparation of (2*E*,4*R**,5*R**)-ethyl-4,5-epoxy-hex-2-enoate (172)

Trifluoroacetic anhydride (14.8 ml, 104 mmol) was added slowly to a suspension of (2*E*,4*E*) ethyl hexa-2,4-dienoate 171 (2.44 g, 17.4 mmol), … provided the epoxide 172 (1.09 g, 7 mmol, 41%) as a colourless oil; **ν_{max} (film)/cm^{-1}:** 2981, 1716 (C=O), 1655 (C=C), 1446, 1378, 1367, 1340, 1302, 1258, 1185, 1140, 1096, 1031, 1005, 975; δ_H **…** Found: [*M*Na]$^+$, 179.060. $[C_8H_{12}O_3Na]^+$ requires 179.0684. Data was consistent with those reported in the literature.[16]

Automatic!



```
<?xml version="1.0" ?>
<cml xmlns="http://www.xml-cml.org/schema">
  <molecule id="m1">
    <name dictRef="nameDict:iupac">
        acetic acid
    </name>
    <name dictRef="nameDict:trivial">
        acetyl hydroxide
    </name>
    <formula inline="AcOH"
        concise="C 2 H 4 O 2"/>
    <atomArray>
      <atom id="a1" elementType="C"
        x2="-2.914" y2="0.769" />
      ...
      <atom id="a8" elementType="H"
        x2="1.086" y2="1.539" />
    </atomArray>
    <bondArray>
      <bond id="b1" atomRefs2="a1 a2"
        order="1" />
      ...
      <bond id="b7" atomRefs2="a3 a8"
        order="1" />
    </bondArray>
  </molecule>
</cml>
```

Human-friendly

Machine-friendly

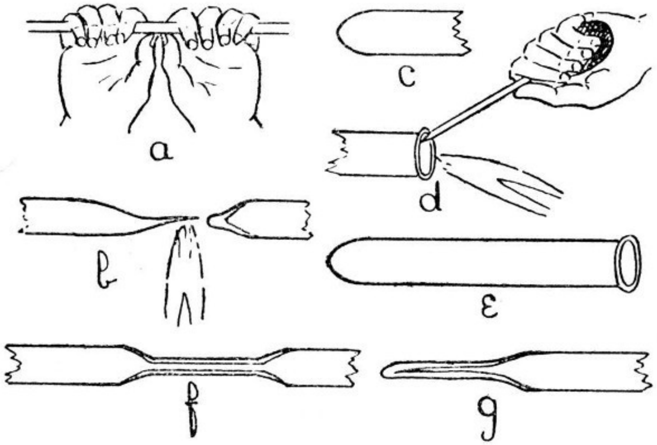
Chemical semantic web (2007)

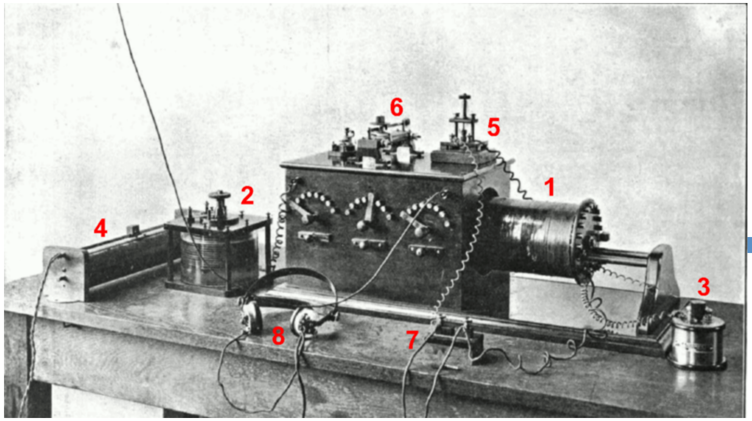
- maybe when you've got an overlay of scalable vector graphics [**CML, InChI and chemical ontologies - everything well-defined and marked up**] on Web 2.0 and access to a semantic Web integrated across a huge space of data, … *Peter Murray-Rust (2007)*

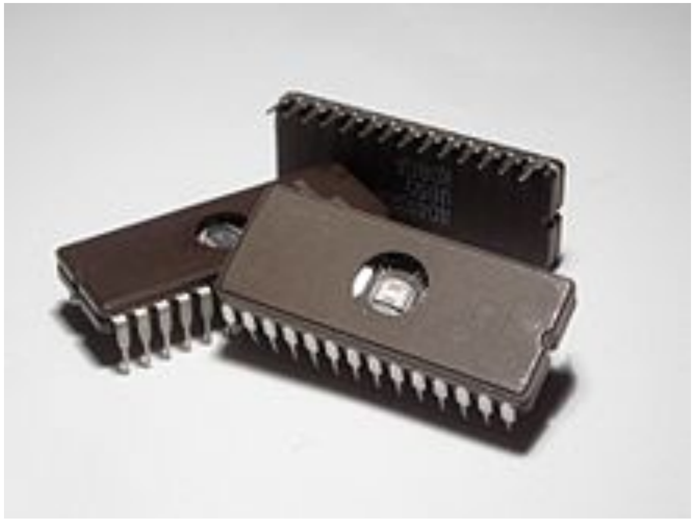
Benefits of semantics

- *"The bane of my life is doing things I know computers could do for me" (Dan Connolly, W3C)*
- Automation
- Reliability
- Interoperability
- Validation
- Transparency
- The semantic web is having a personal amanuensis

Componentised approach liberates





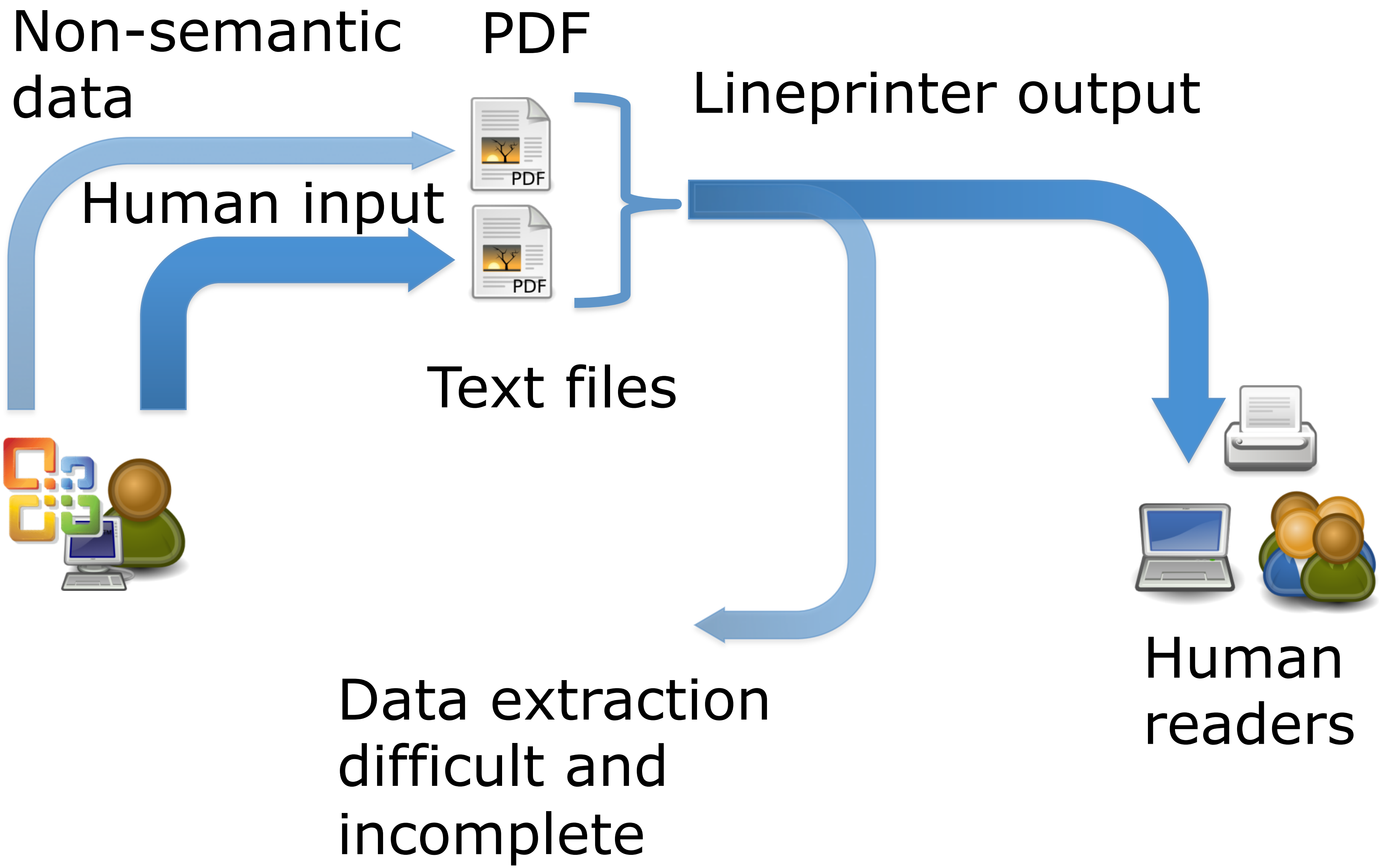


Individual, manual,
unreusable, flaky

Commodity, standard,
reliable, re-usable

Current scientific information flow

… is broken for data-rich science



Non-semantic data

PDF

Lineprinter output

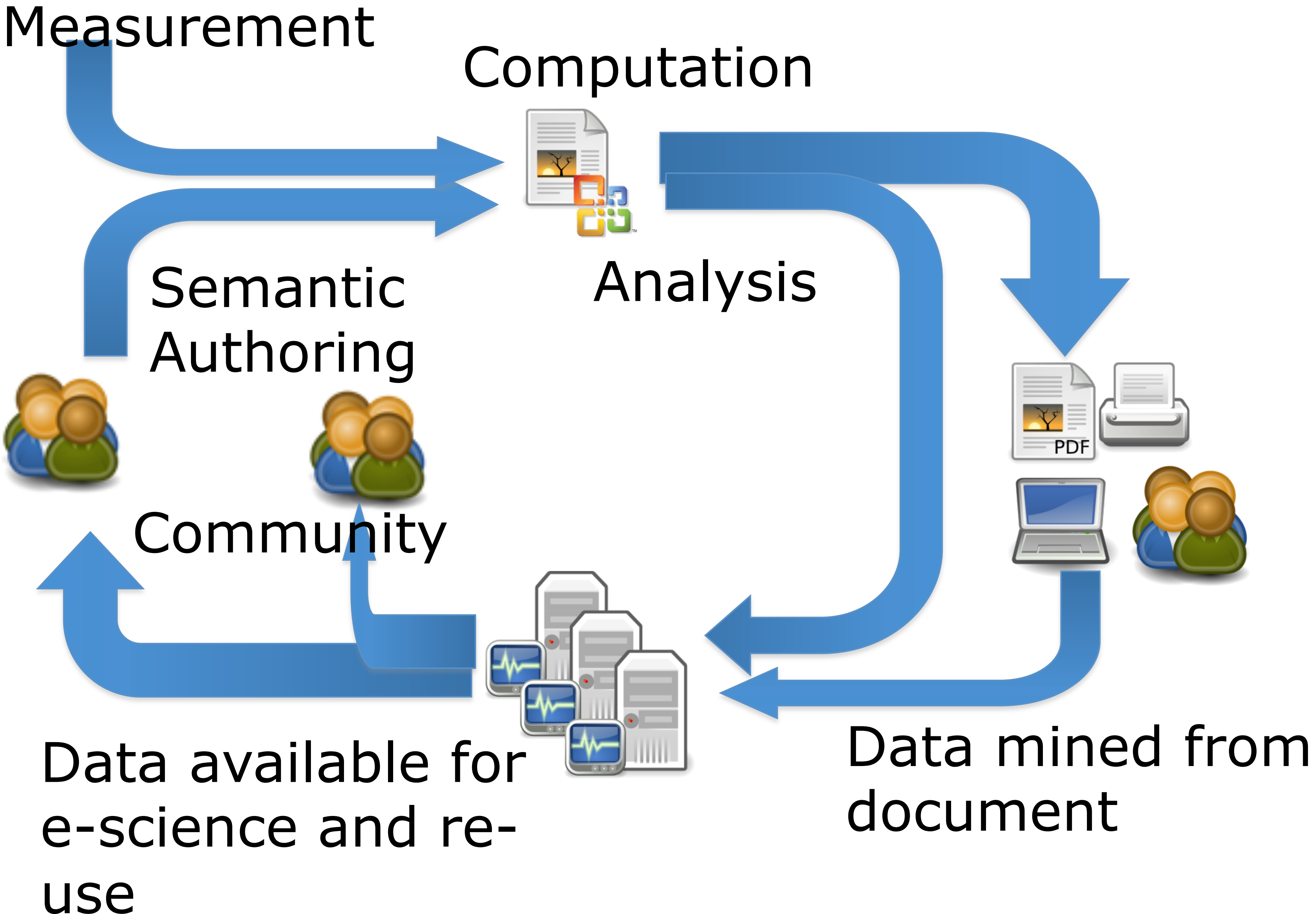
Human input

Text files

Data extraction difficult and incomplete

Human readers

Semantic network closes the loop



Measurement

Computation

Semantic Authoring

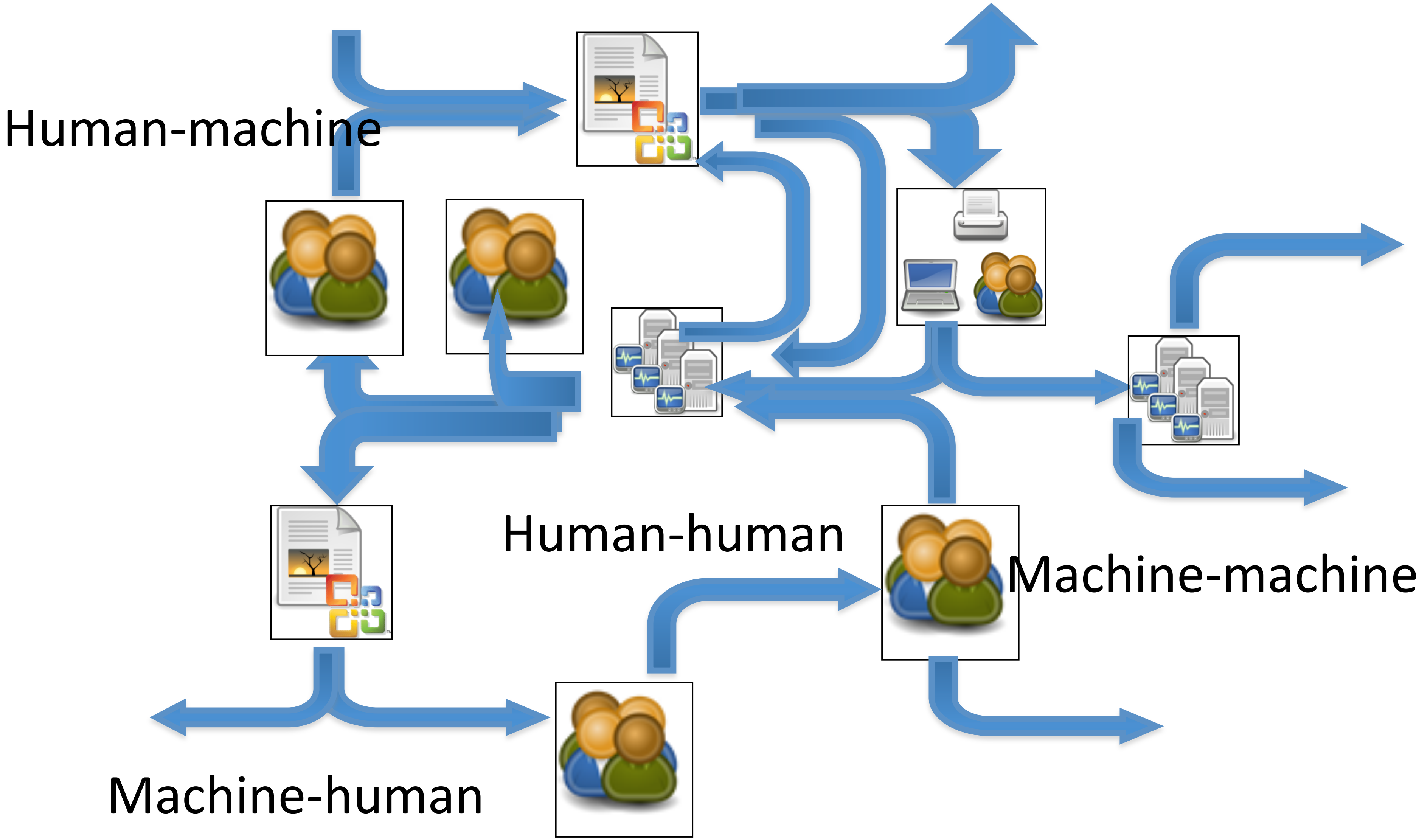
Analysis

Community

Data available for e-science and re-use

Data mined from document

The network grows autonomously



Human-machine

Human-human

Machine-machine

Machine-human

Overview

- Semantics: Telling machines PRECISELY what we mean
- Translating machine output into human language
- Extracting/translating our current chemistry into semantic form:
    1. Program output
    2. Chemical databases
    3. Natural Language Processing (written) (NLP)

Representing Semantics

Complementary approaches:

Markup Languages ("hardcoded objects") MathML, G(eo)ML, CellML, S(ys)B(io)ML,

- CML (Chemistry and numeric science):
  1. Molecules (atoms, bonds, coordinates,
  2. Reactions,
  3. Spectra,
  4. Solid state,
  5. Computation

RDF (relationships, annotations, linking).

Problem: Explaining chemistry to a machine

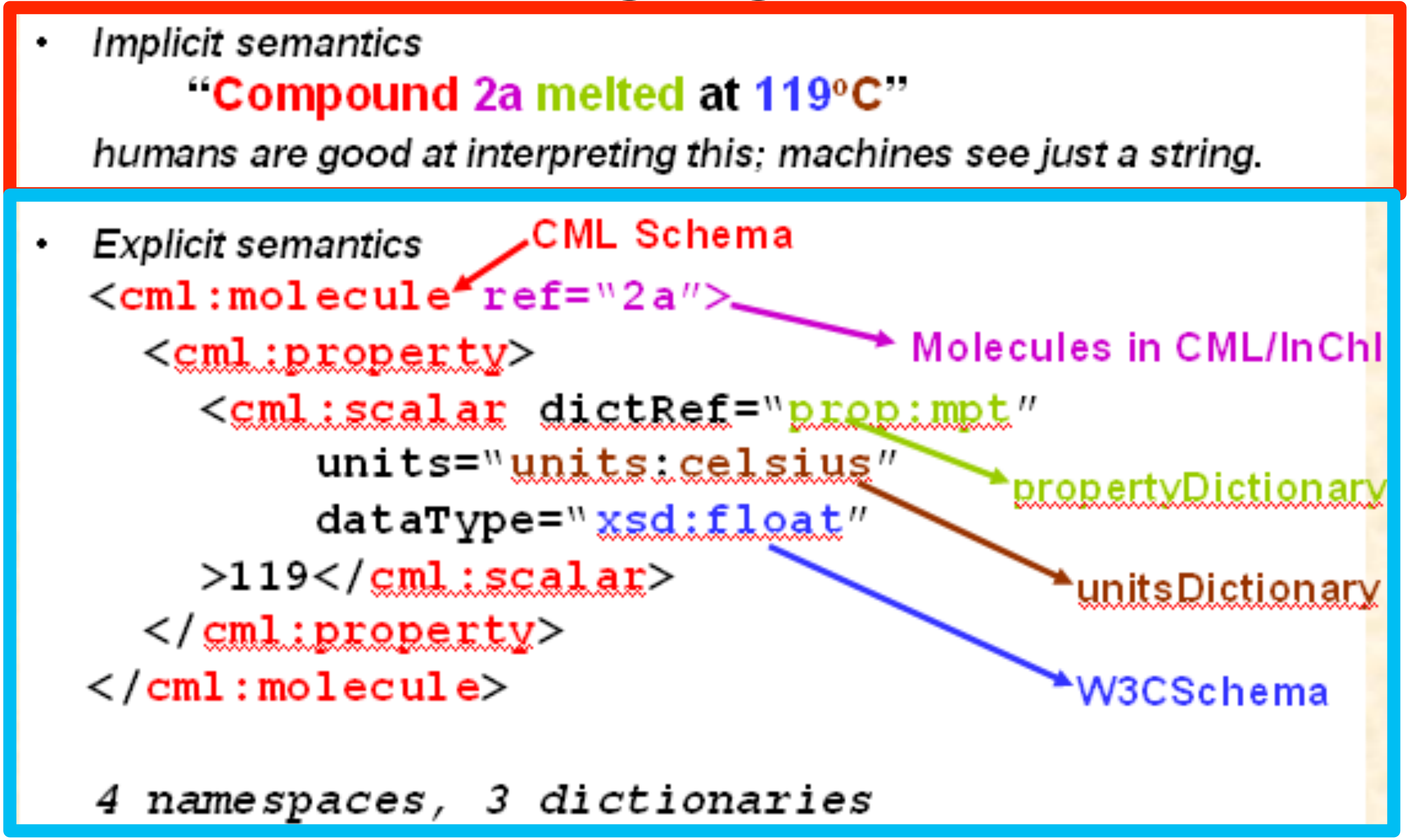
“The calculated dipole moment of ethanol was 1.6 D”

The machine “understands” basic chemical structure (atoms, bonds and coordinates) and numeric properties (“1.6”). It does not understand

- “calculated”,
- “dipole moment”,
- “ethanol”,
- “D”.

NOW: Communal discussion of how to tackle this

Humans and machines use different languages



- *Implicit semantics*

  "Compound 2a melted at 119°C"

  humans are good at interpreting this; machines see just a string.

- *Explicit semantics*

```
<cml:molecule ref="2a">
  <cml:property>
    <cml:scalar dictRef="prop:mpt"
          units="units:celsius"
          dataType="xsd:float"
    >119</cml:scalar>
  </cml:property>
</cml:molecule>
```

CML Schema

Molecules in CML/InChI

propertyDictionary

unitsDictionary

W3CSchema

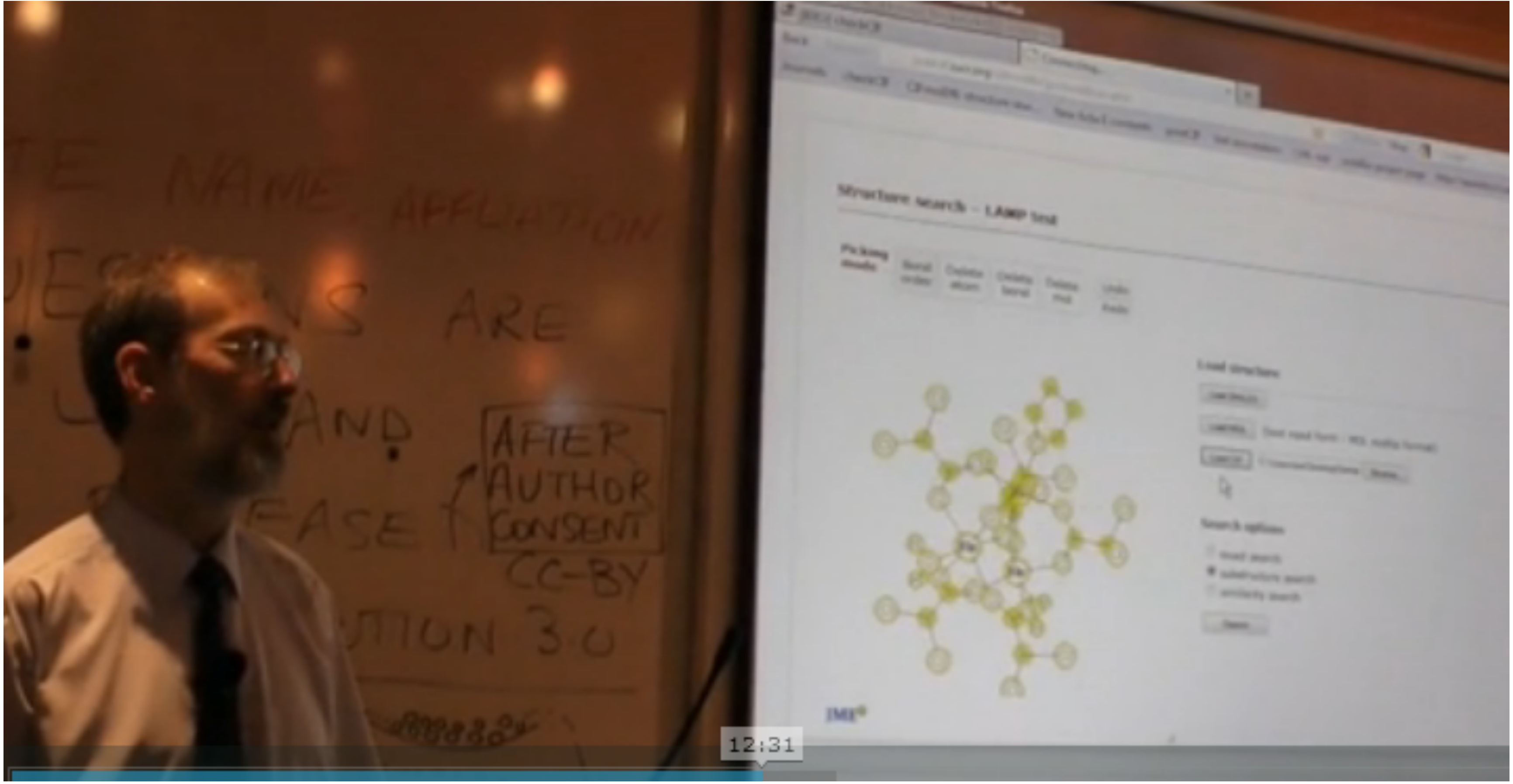
4 namespaces, 3 dictionaries

DIPOLE MOMENT OF ETHANOL

```
<cml:molecule xmlns:cml=“http://www.xml-cml.org/schema” “ title=“ethanol”>
  <cml:atomArray>
    <cml:atom id=“a1” elementType=“O” x3=“0.0”, y3=“0.0, z3=“0.0”/>
    <cml:atom id=“a2” elementType=“C” x3=“0.0”, y3=“0.0, z3=“1.54”/>
    <cml:atom id=“a3” elementType=“C” x3=“0.0”, y3=“1.2”, z3=“2.2”/>
        <!- atoms omitted →
    <cml:atom id=“a1h” elementType=“H” x3=“0.0”, y3=“-0.8”, z3=“-0.4”/>
  </cml:atomArray>
  <cml:bondArray>
    <cml:bond id=“a1_a2” atomRefs2=“a1 a2” order=“S”/>
    <cml:bond id=“a2_a3” atomRefs2=“a2 a3” order=“S”/>
    <!- bonds omitted →
  </cml:bondArray>
  <cml:property dictRef=“compchem:scalarDipole” role=“compchem:calculated”>
    <cml:scalar dataType=“xsd:double”
         units=“compchem:debye”>1.60</cml:scalar>
  </cml:property>
  <cml:property dictRef=“compchem:vectorDipole” role=“compchem:calculated”>
    <cml:vector3
         units=“compchem:debye”>1.1 1.3 0.2</cml:vector3>
  </cml:property>
</cml:molecule>
```

Semantic authoring IUCr

- http://blogs.ch.cam.ac.uk/pmr/2012/01/23/brian-mcmahon-publishing-semantic-crystallography-every-science-data-publisher-should-watch-this-all-the-way-through/



Sociopolitical aspects

- Little communal interest in formalising chemistry (exceptions: InChI, IUPAC books)
- Most initiatives are bottom-up (CML, Computational Materials, PubChem, Wikipedia)
- Broken publication system (no semantics and widespread legal prohibition of machine extraction from literature)

Challenge

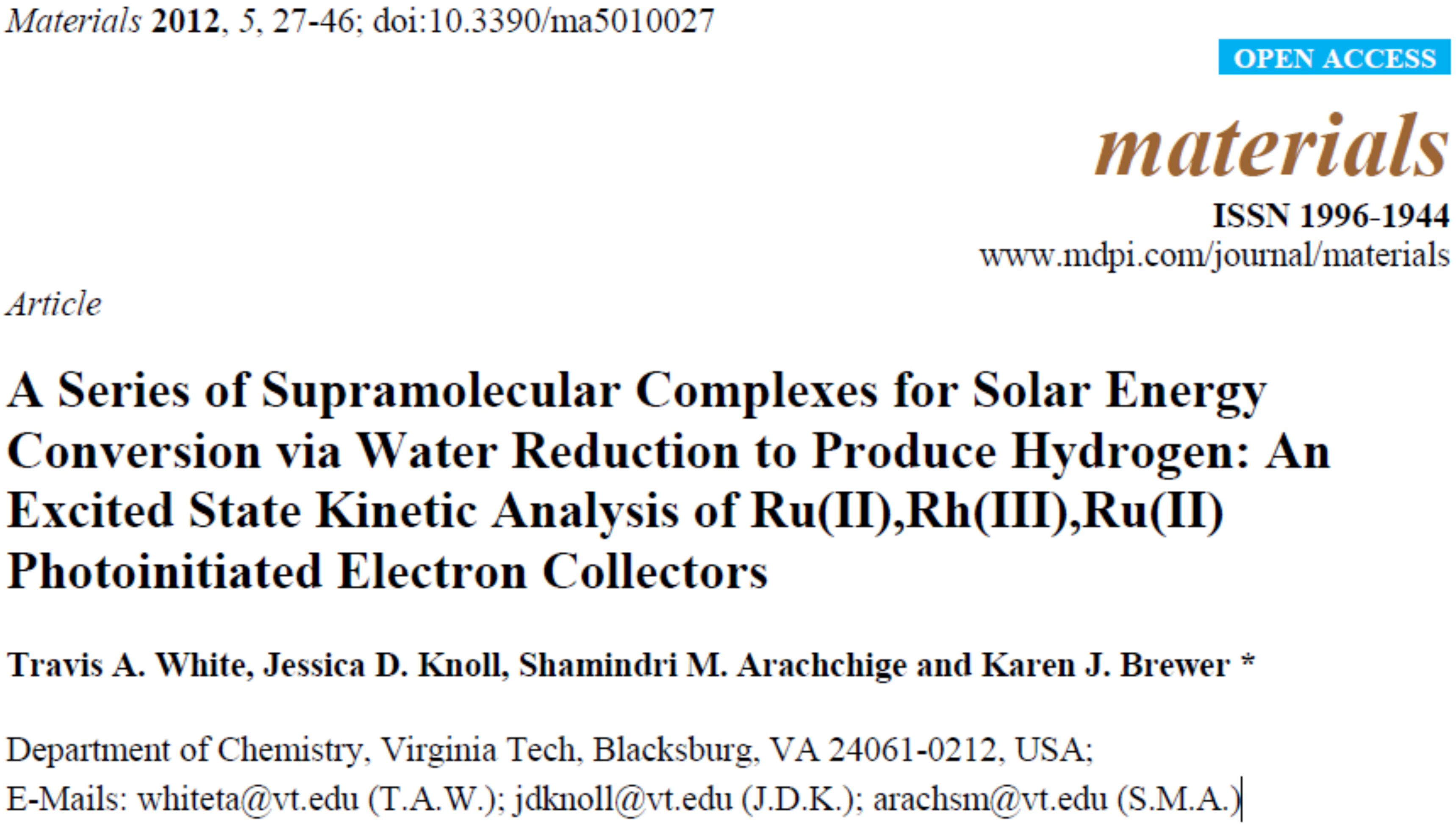
Extracting semantic information from a typical materials paper.

http://www.mdpi.com/1996-1944/5/1/27

Could you reproduce this work?

Could you use the data?

- **Example**: *Materials* **2012**, *5*, 27-46; doi:



Materials **2012**, *5*, 27-46; doi:10.3390/ma5010027

OPEN ACCESS

materials

ISSN 1996-1944
www.mdpi.com/journal/materials

Article

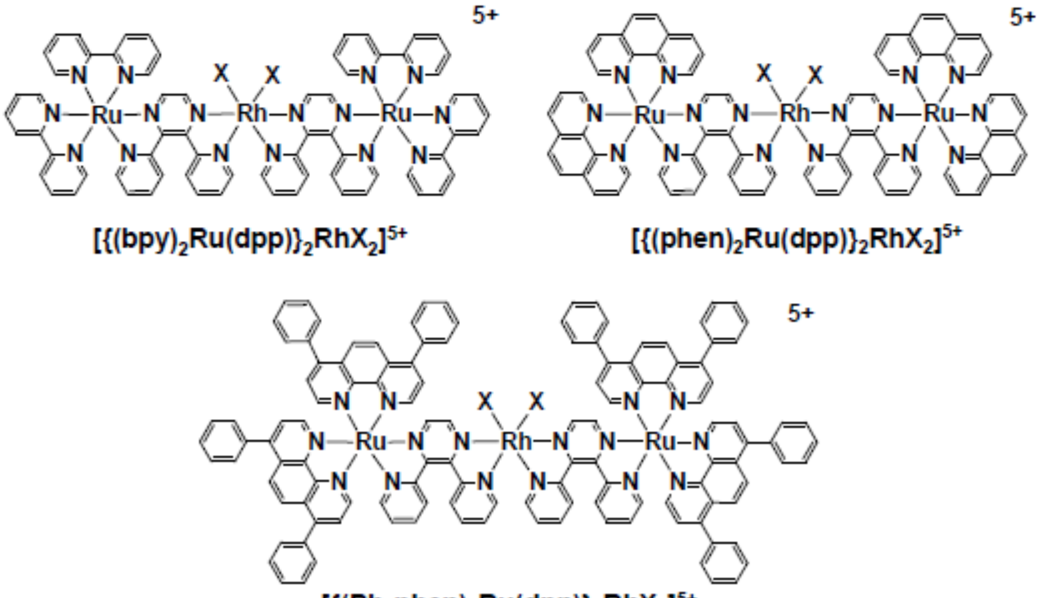
A Series of Supramolecular Complexes for Solar Energy Conversion via Water Reduction to Produce Hydrogen: An Excited State Kinetic Analysis of Ru(II),Rh(III),Ru(II) Photoinitiated Electron Collectors

Travis A. White, Jessica D. Knoll, Shamindri M. Arachchige and Karen J. Brewer *

Department of Chemistry, Virginia Tech, Blacksburg, VA 24061-0212, USA;
E-Mails: whiteta@vt.edu (T.A.W.); jdknoll@vt.edu (J.D.K.); arachsm@vt.edu (S.M.A.)

CHEMICAL STRUCTURES

Figure 2. Ru(II),Rh(III),Ru(II) photoinitiated electron collectors of the supramolecular architecture $[\{(TL)_2Ru(dpp)\}_2RhX_2]^{5+}$ (TL = bpy = 2,2′-bipyridine, phen = 1,10-phenanthroline, Ph_2phen = 4,7-diphenyl-1,10-phenanthroline; dpp = 2,3-bis (2-pyridyl)pyrazine; X = Cl or Br).

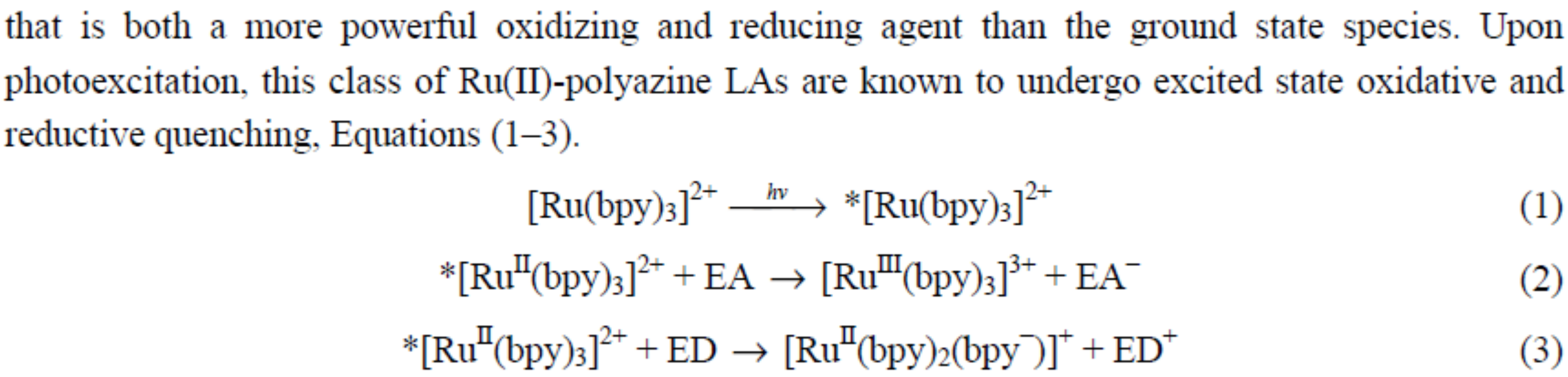


$[\{(bpy)_2Ru(dpp)\}_2RhX_2]^{5+}$

$[\{(phen)_2Ru(dpp)\}_2RhX_2]^{5+}$

$[\{(Ph_2phen)_2Ru(dpp)\}_2RhX_2]^{5+}$

REACTIONS



that is both a more powerful oxidizing and reducing agent than the ground state species. Upon photoexcitation, this class of Ru(II)-polyazine LAs are known to undergo excited state oxidative and reductive quenching, Equations (1–3).

$$[\mathrm{Ru(bpy)_3}]^{2+} \xrightarrow{h\nu} {}^*[\mathrm{Ru(bpy)_3}]^{2+} \quad (1)$$

$$^*[\mathrm{Ru^{II}(bpy)_3}]^{2+} + \mathrm{EA} \rightarrow [\mathrm{Ru^{III}(bpy)_3}]^{3+} + \mathrm{EA}^- \quad (2)$$

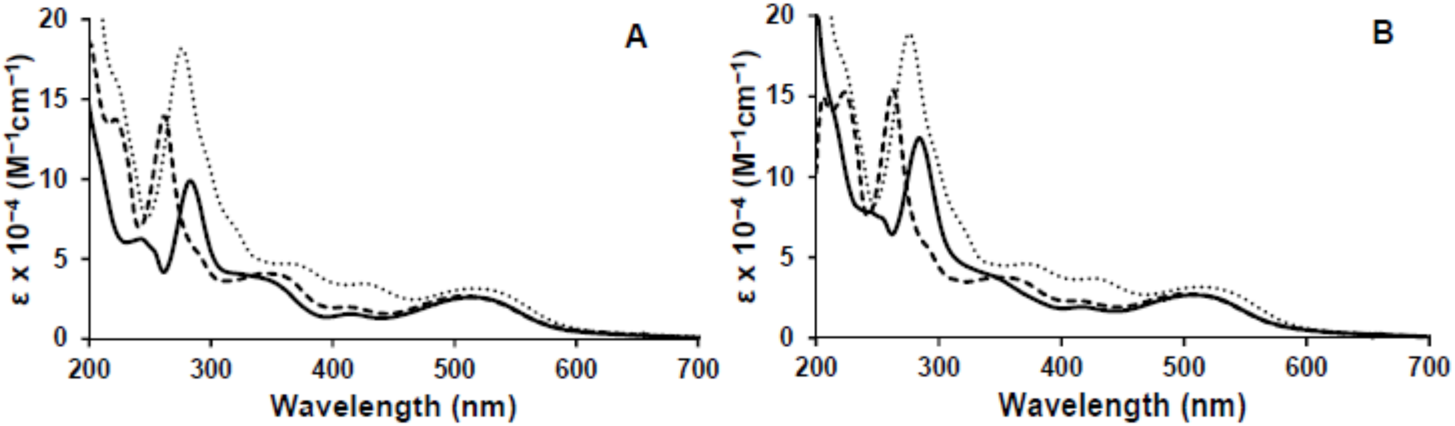
$$^*[\mathrm{Ru^{II}(bpy)_3}]^{2+} + \mathrm{ED} \rightarrow [\mathrm{Ru^{II}(bpy)_2(bpy^-)}]^{+} + \mathrm{ED}^+ \quad (3)$$

ABBREVIATIONS

“… electron donor (ED), such as an electron rich, metal-based light absorber (LA), and electron acceptor (EA) sites.”

SPECTRA

Figure 3. Electronic absorption spectra for the complexes (**A**) $[\{(TL)_2Ru(dpp)\}_2RhCl_2]^{5+}$, where TL = bpy (——), phen (- - -), Ph_2phen (· · ·) and (**B**) $[\{(TL)_2Ru(dpp)\}_2RhBr_2]^{5+}$, where TL = bpy (——), phen (- - -), Ph_2phen (· · ·).



TABLES

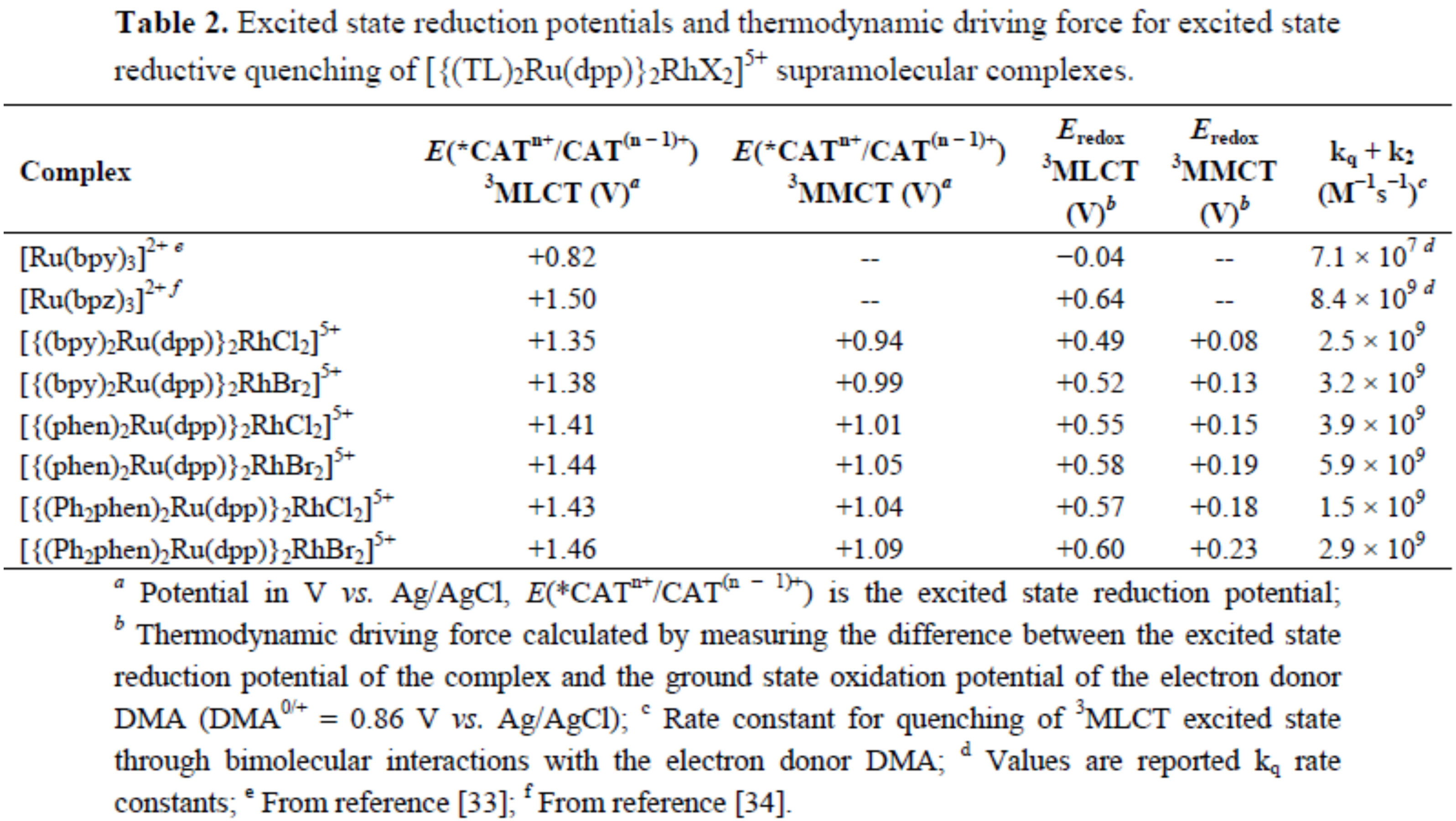
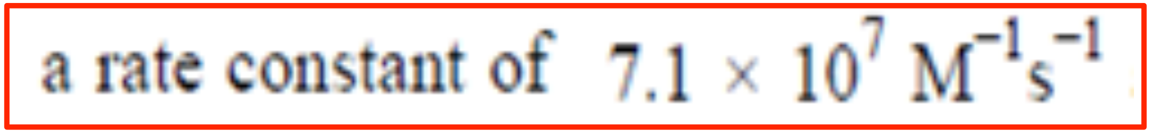


Table 2. Excited state reduction potentials and thermodynamic driving force for excited state reductive quenching of $[\{(TL)_2Ru(dpp)\}_2RhX_2]^{5+}$ supramolecular complexes.

Complex	$E(^*CAT^{n+}/CAT^{(n-1)+})$ ^{3}MLCT (V)^a	$E(^*CAT^{n+}/CAT^{(n-1)+})$ ^{3}MMCT (V)^a	E_{redox} ^{3}MLCT (V)^b	E_{redox} ^{3}MMCT (V)^b	$k_q + k_2$ ($M^{-1}s^{-1}$)^c
$[Ru(bpy)_3]^{2+\,e}$	+0.82	--	−0.04	--	$7.1 \times 10^{7\,d}$
$[Ru(bpz)_3]^{2+\,f}$	+1.50	--	+0.64	--	$8.4 \times 10^{9\,d}$
$[\{(bpy)_2Ru(dpp)\}_2RhCl_2]^{5+}$	+1.35	+0.94	+0.49	+0.08	2.5×10^9
$[\{(bpy)_2Ru(dpp)\}_2RhBr_2]^{5+}$	+1.38	+0.99	+0.52	+0.13	3.2×10^9
$[\{(phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$	+1.41	+1.01	+0.55	+0.15	3.9×10^9
$[\{(phen)_2Ru(dpp)\}_2RhBr_2]^{5+}$	+1.44	+1.05	+0.58	+0.19	5.9×10^9
$[\{(Ph_2phen)_2Ru(dpp)\}_2RhCl_2]^{5+}$	+1.43	+1.04	+0.57	+0.18	1.5×10^9
$[\{(Ph_2phen)_2Ru(dpp)\}_2RhBr_2]^{5+}$	+1.46	+1.09	+0.60	+0.23	2.9×10^9

[a] Potential in V *vs.* Ag/AgCl, $E(^*CAT^{n+}/CAT^{(n-1)+})$ is the excited state reduction potential; [b] Thermodynamic driving force calculated by measuring the difference between the excited state reduction potential of the complex and the ground state oxidation potential of the electron donor DMA ($DMA^{0/+}$ = 0.86 V *vs.* Ag/AgCl); [c] Rate constant for quenching of ^{3}MLCT excited state through bimolecular interactions with the electron donor DMA; [d] Values are reported k_q rate constants; [e] From reference [33]; [f] From reference [34].

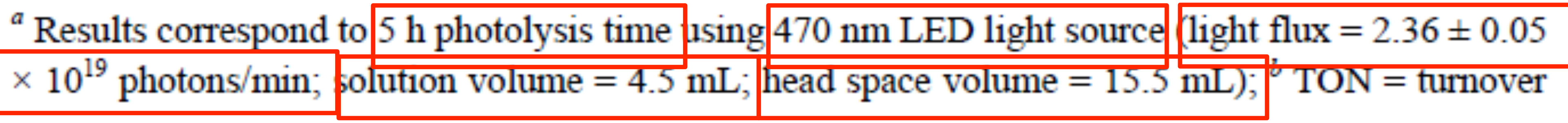
PROPERTIES (NAME-VALUE-UNITS)



a rate constant of 7.1×10^{7} $M^{-1}s^{-1}$

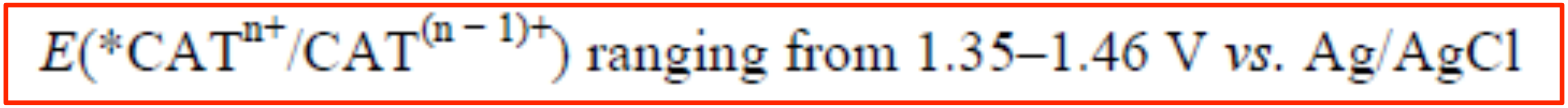
Name Value Units

V U N V U N N V V



[a] Results correspond to 5 h photolysis time using 470 nm LED light source (light flux = 2.36 ± 0.05 × 10^{19} photons/min; solution volume = 4.5 mL; head space volume = 15.5 mL); [b] TON = turnover

U



E(*CAT^{n+}/CAT$^{(n-1)+}$) ranging from 1.35–1.46 V *vs.* Ag/AgCl

N V V U

Note CML supports value ranges and errors

Mathematics

$$\frac{1}{\Phi_{\mathrm{product}}} = \left(\frac{1}{\Phi_{3_{\mathrm{MMCT}}}}\right)\left(\frac{\mathrm{k}_4}{\mathrm{k}_{\mathrm{q2}}[\mathrm{DMA}]}\right) + \frac{\mathrm{k}_{\mathrm{q2}} + \mathrm{k}_3}{\mathrm{k}_{\mathrm{q2}}}$$

CML is being integrated with computable (content) MathML

Materials Search Challenge

- What would you like a “Google for materials” to find for you in the scientific literature?

Creating CML

- ~~Hand-editing (tedious and errorprone)~~
- Tools (Avogadro, JChempaint, Chem4Word)
- Direct output from programs (FoX, JUMBO)
- Conversion from structured files (Openbabel)
- Online knowledgebase (Wikipedia, PubChem)
- Conversion from semistructured (log)files (JUMBOConverters)
- Extraction from text (ChemicalTagger, OSCAR, OPSIN, AMI2-SVG2CML)

Demos

- OPSIN http://opsin.ch.cam.ac.uk
- ChemicalTagger http://chemicaltagger.ch.cam.ac.uk

Crystaleye

- A database of 200,000 crystal structures scraped from supplemental information
- CML molecules and name-value pairs
- Re-usable as fragment base

http://wwmm.ch.cam.ac.uk/crystaleye

Knowledgebases

- Quixote. Logfiles from compchem output parsed into CML
- Integrated into an XML/RDF knowledgebase
- Searchable on chemistry and properties
- http://quixote.ch.cam.ac.uk
- Sam Adams, Cambridge

Ontologies in physical science

- #animalgarden production

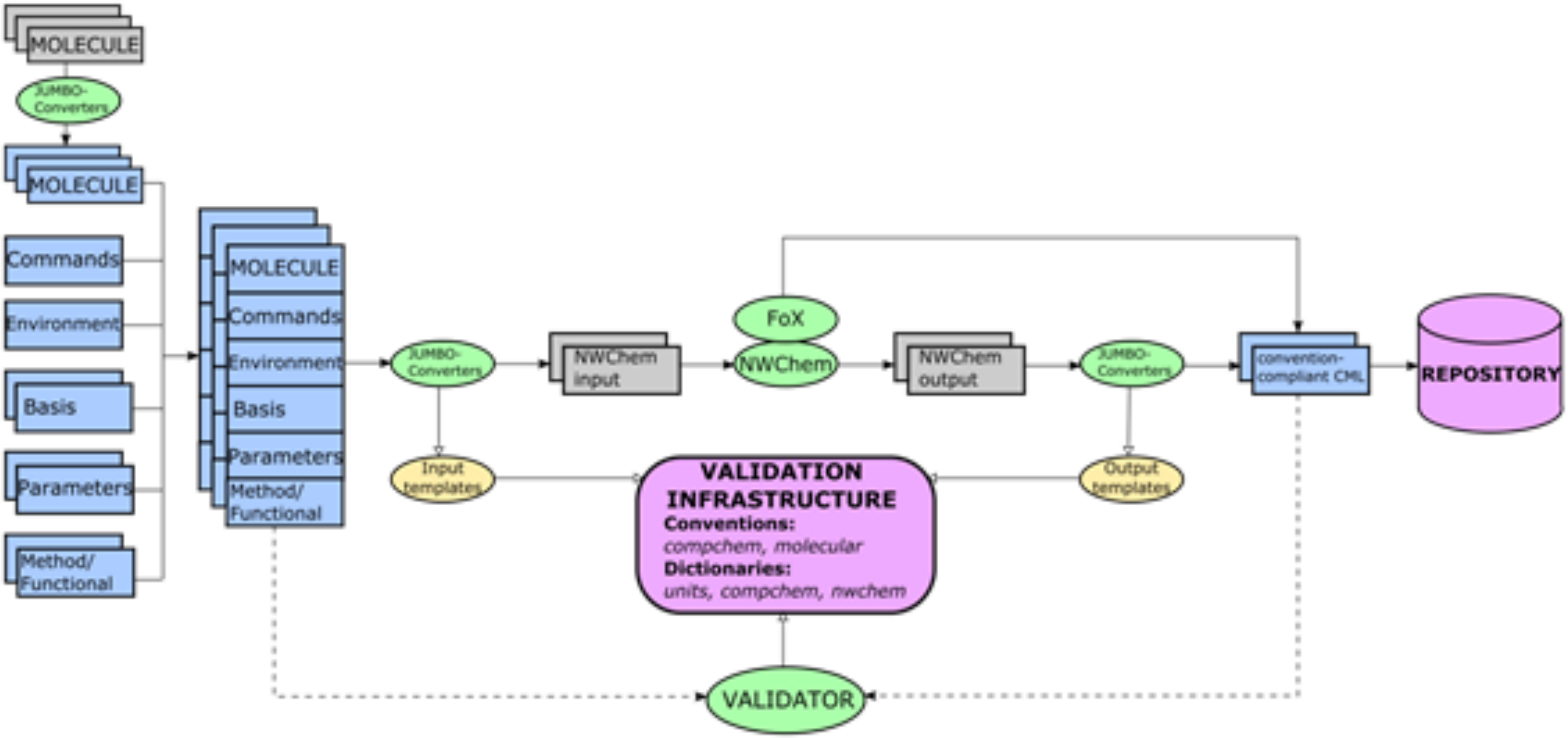


FIGURE 2. Schematic model of a semantic framework for computational chemistry (using NWChem[11]

Jumbo Converters

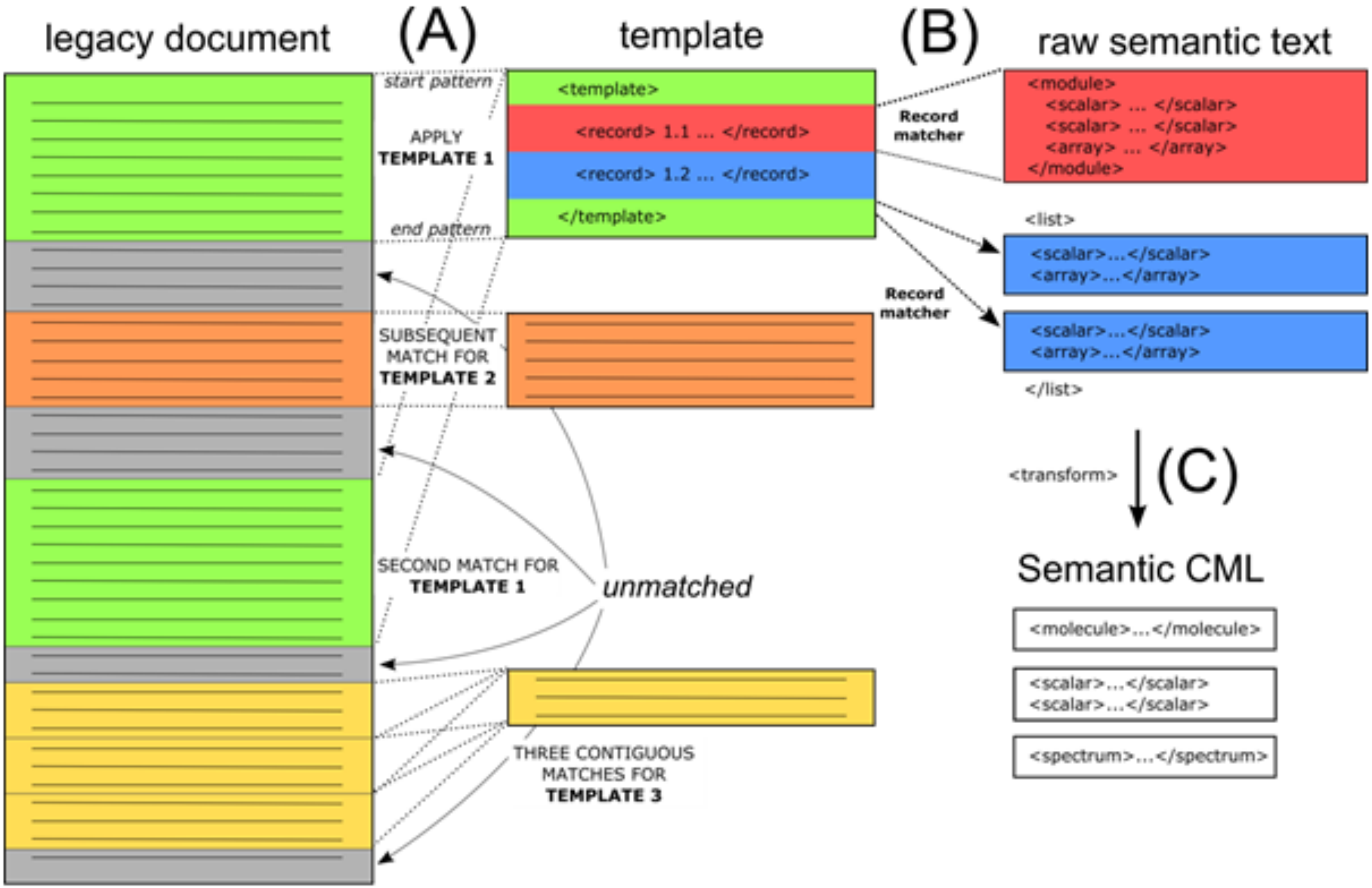
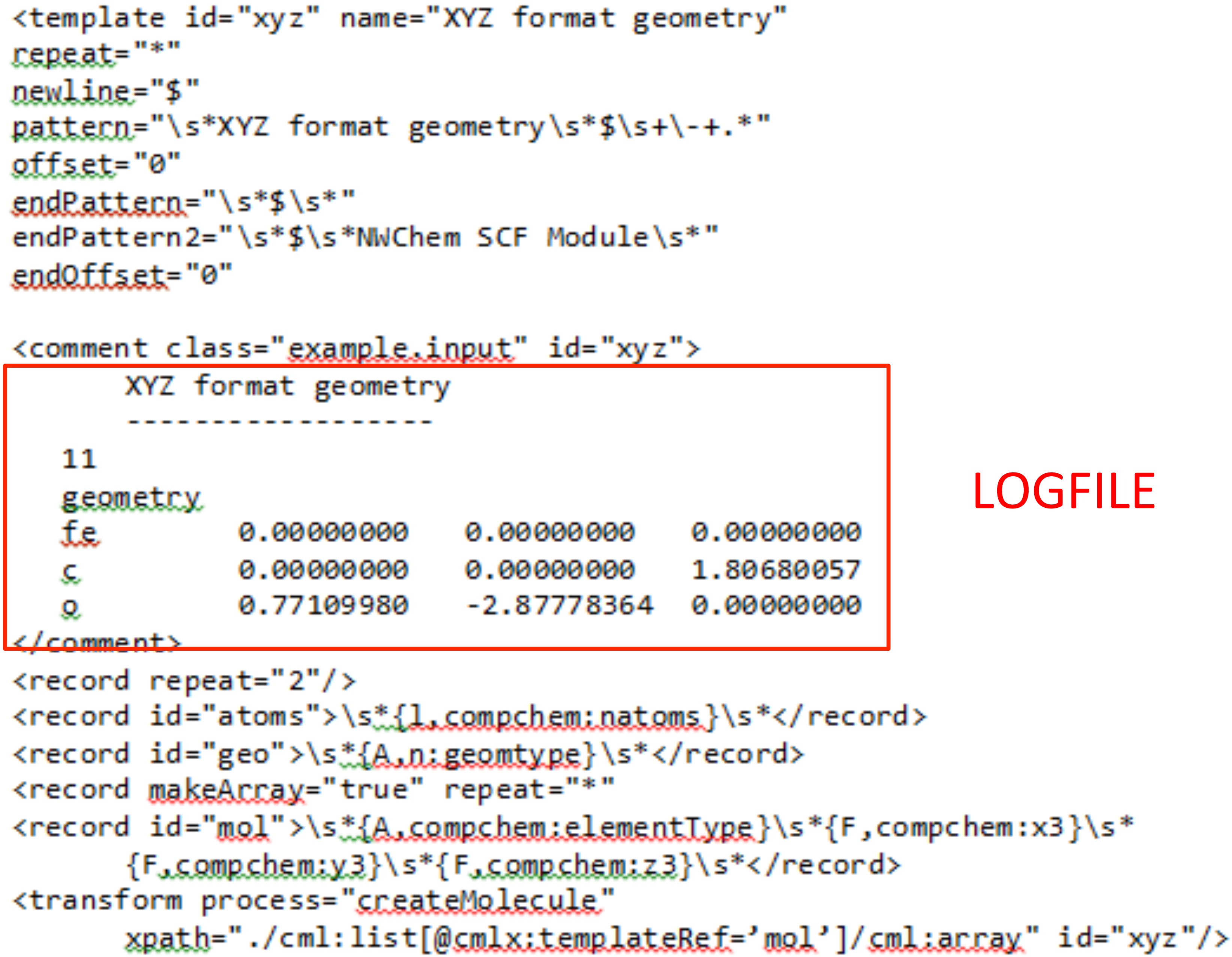


FIGURE 3: Processing a legacy document with templates. (A) A template matches chunks of the

JumboConverters - templates



```
<template id="xyz" name="XYZ format geometry"
repeat="*"
newline="$"
pattern="\s*XYZ format geometry\s*$\s+\-+.*"
offset="0"
endPattern="\s*$\s*"
endPattern2="\s*$\s*NWChem SCF Module\s*"
endOffset="0"

<comment class="example.input" id="xyz">
      XYZ format geometry
      -------------------
   11
   geometry
   fe        0.00000000    0.00000000    0.00000000
   c         0.00000000    0.00000000    1.80680057
   o         0.77109980   -2.87778364    0.00000000
</comment>
<record repeat="2"/>
<record id="atoms">\s*{I,compchem:natoms}\s*</record>
<record id="geo">\s*{A,n:geomtype}\s*</record>
<record makeArray="true" repeat="*"
<record id="mol">\s*{A,compchem:elementType}\s*{F,compchem:x3}\s*
      {F,compchem:y3}\s*{F,compchem:z3}\s*</record>
<transform process="createMolecule"
      xpath="./cml:list[@cmlx:templateRef=’mol’]/cml:array" id="xyz"/>
```

LOGFILE

Dictionaries

- http://www.xml-cml.org/convention/unit-dictionary
- http://www.xml-cml.org/convention/compchem

JumboConverters Structure

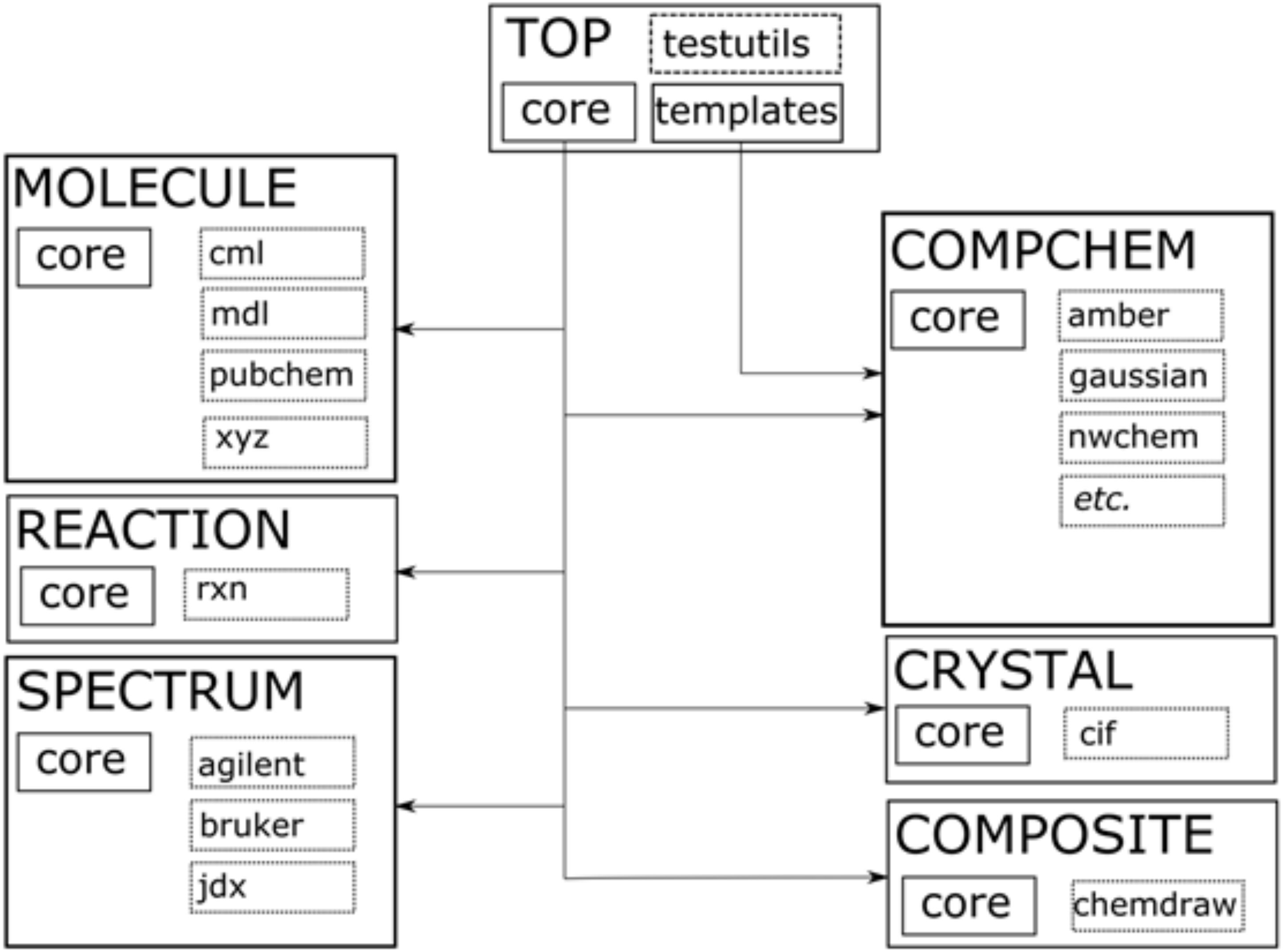


Figure 6: The modular structure of JUMBO-Converters. The five subdomains of chemistry are

- http://pantonprinciples.org/ Open data

Research data: managing and training

- http://blogs.ch.cam.ac.uk/pmr/2013/02/13/rds2013-principles-for-managing-research-data/
- http://sophiekershaw.wordpress.com/author/sophiekershaw/
- http://www.opensciencetraining.com/content.php

TimBl’s Open data
http://5stardata.info

★ make your stuff available on the Web (whatever format) under an open license

★★ make it available as structured data (e.g., Excel instead of image scan of a table)

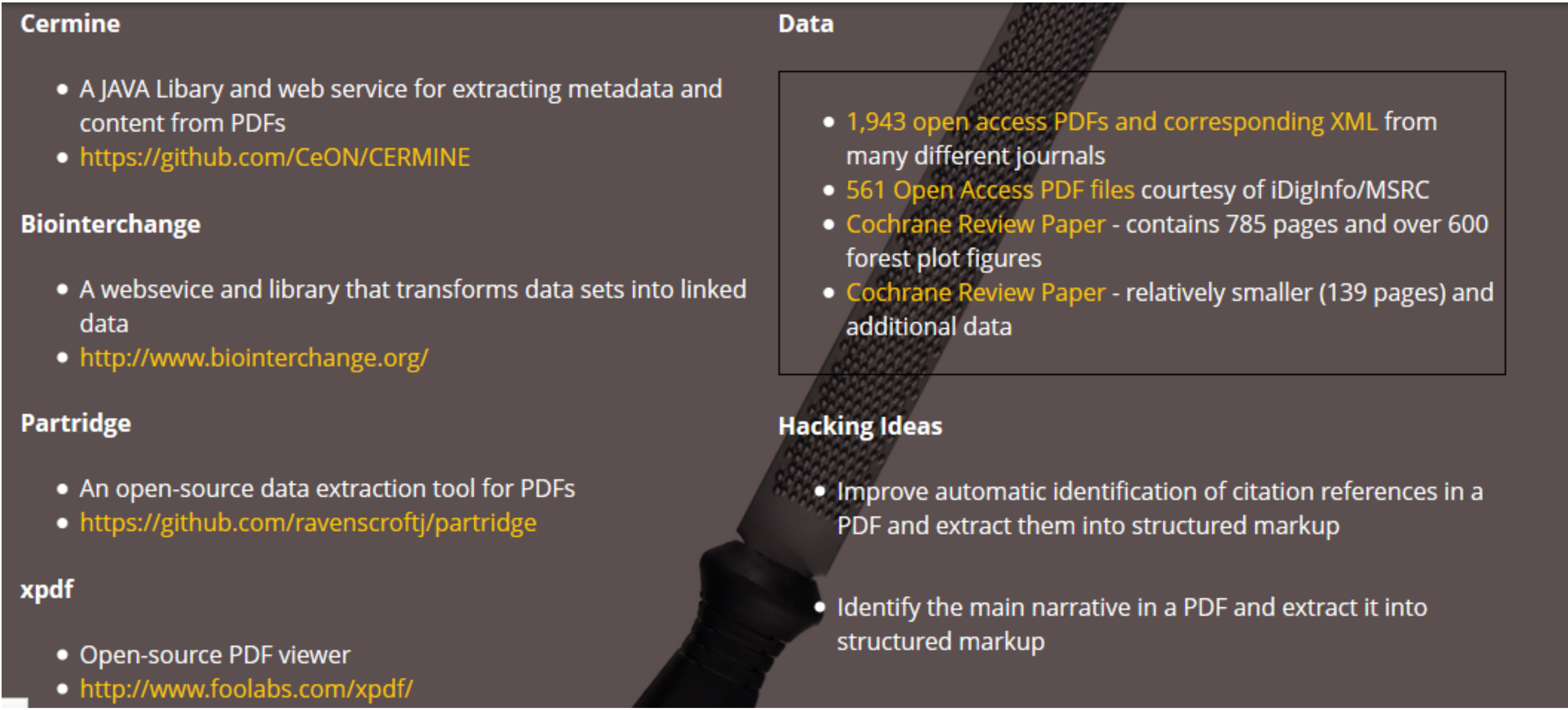
★★★ use non-proprietary formats (e.g., CSV instead of Excel)

★★★★ use URIs to denote things, so that people can point at your stuff

★★★★★ link your data to other data to provide context

Jailbreaking the PDF Hackathon

- http://scholrev.org/hackathon/ a group of enthusiasts committed to liberating data.



Cermine

- A JAVA Libary and web service for extracting metadata and content from PDFs
- https://github.com/CeON/CERMINE

Biointerchange

- A websevice and library that transforms data sets into linked data
- http://www.biointerchange.org/

Partridge

- An open-source data extraction tool for PDFs
- https://github.com/ravenscroftj/partridge

xpdf

- Open-source PDF viewer
- http://www.foolabs.com/xpdf/

Data

- 1,943 open access PDFs and corresponding XML from many different journals
- 561 Open Access PDF files courtesy of iDigInfo/MSRC
- Cochrane Review Paper - contains 785 pages and over 600 forest plot figures
- Cochrane Review Paper - relatively smaller (139 pages) and additional data

Hacking Ideas

- Improve automatic identification of citation references in a PDF and extract them into structured markup
- Identify the main narrative in a PDF and extract it into structured markup

Conclusions