Crystallography and the Semantic Web

Peter Murray-Rust
University of Cambridge
Open Knowledge Foundation

IUCr COMCIFs workshop, Warwick, 2013-08-25

Unilever Cambridge
Centre for Molecular Informatics

Open Knowledge Foundation

CC 0 ZERO

Themes

- What is OPEN?
- Governments and Funder mandates for Open Data.
- The Semantic Web.
- Crystallography Open Database and Crystaleye
- Recommendations for Open Crystallography

Funding includes JISC, Unilever, EPSRC.

The inspiration of IUCr

"[we] owe a huge debt to the International Union of Crystallography (IUCr) and Brian McMahon…

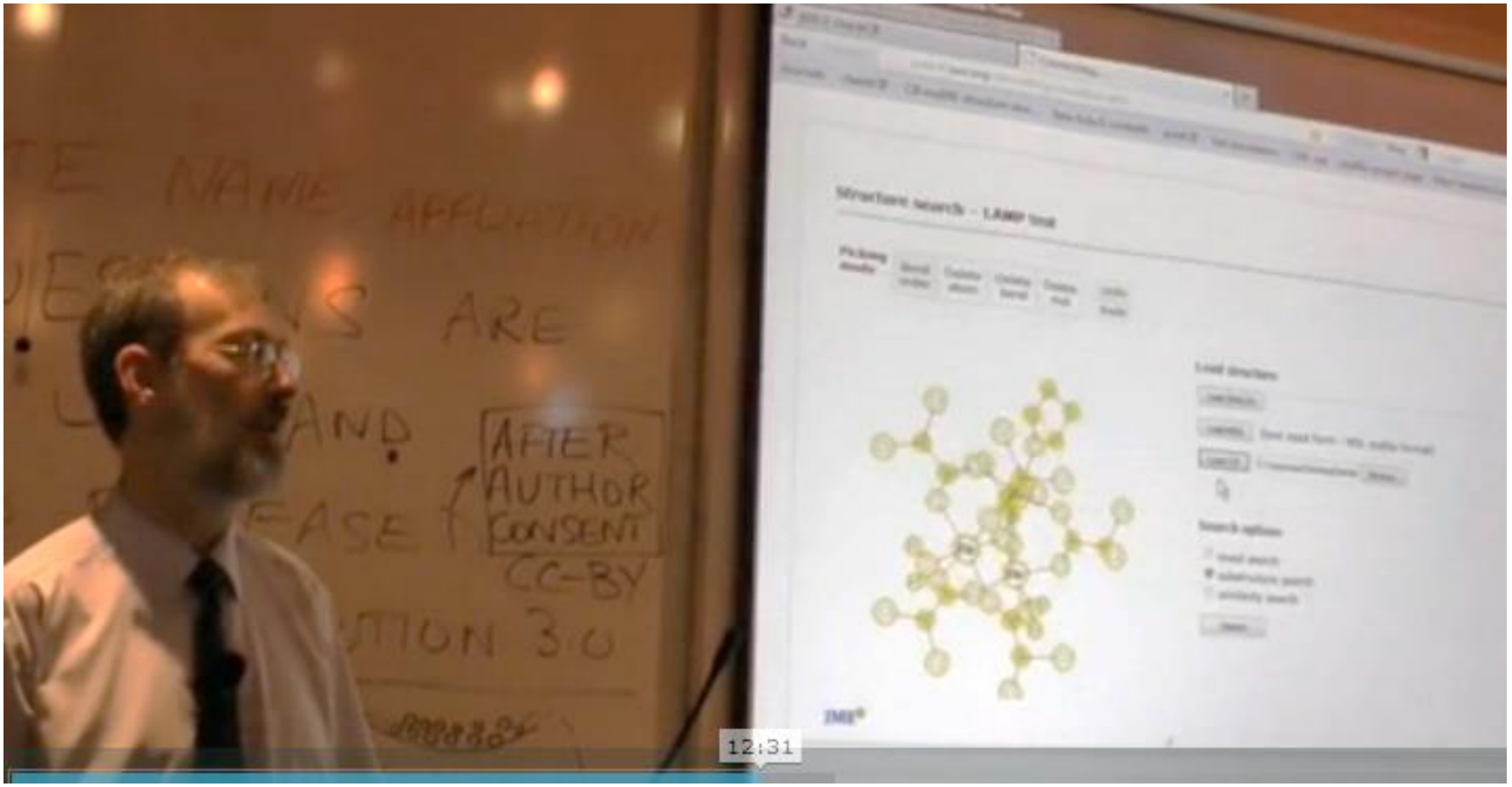
Quite simply they are the best semantic scientific publishers of the current century. They also have the best community-base for scientific publishing that I know. …

…they have steadily invested in the information infrastructure (ontology) of their discipline. And it's been a community effort…

Murray-Rust and Rzepa: Journal of Cheminformatics 2012, 4:14 http://www.jcheminf.com/content/4/1/14

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/



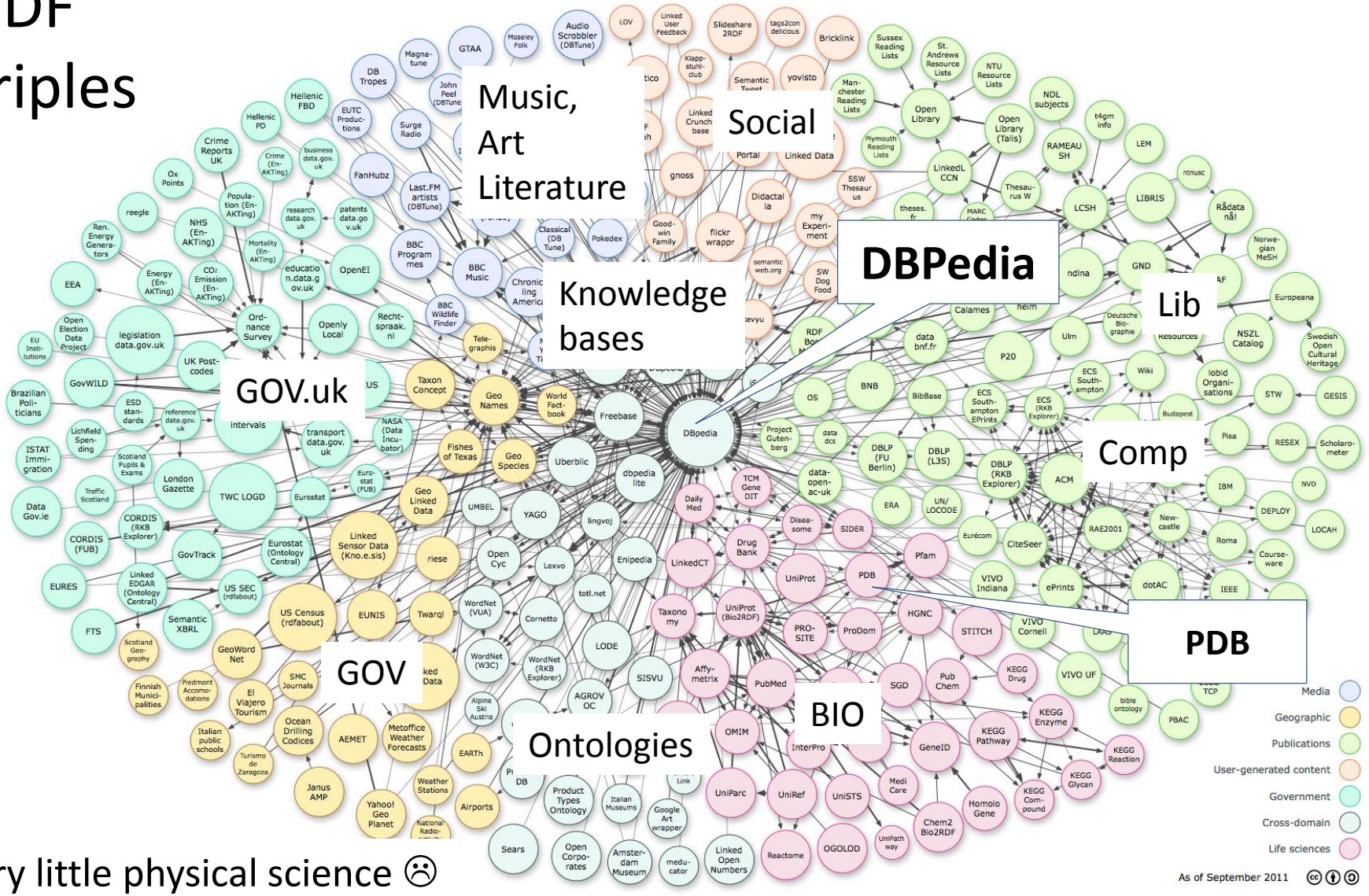
The Semantic Web

"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*

Linked **Open** Data – the world's knowledge

RDF triples



very little physical science ☹

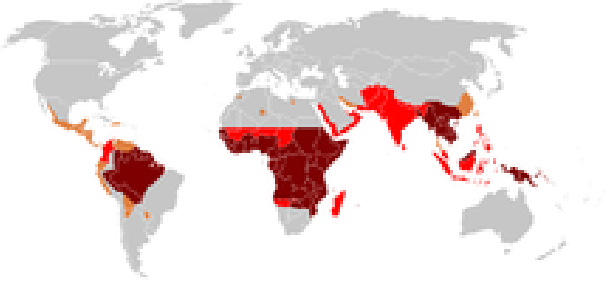
http://upload.wikimedia.org/wikipedia/commons/3/34/LOD_Cloud_Diagram_as_of_September_2011.png

Linked Open data from Wikipedia

"Which Rivers flow into the Rhine and are longer than 50 kilometers?" or "Which Skyscrapers in China have more than 50 floors and have been constructed before the year 2000?"

Open Crystallography?

"Which countries where tropical diseases are endemic have published structures of chiral natural products?"



CC-BY-SA from Wikipedia

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 crystal structures** from them
- Compute **physical properties** with **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

Semantics for Crystal Materials Informatics

- 1993 COMCIFS
- 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, ORESChem (Microsoft Research)
- 2011 PNNL (US) meetings and visit
- 2012 Semantic Physical Science (Cambridge)
- 2013 CSIRO Materials Science Informatics

Open Definition

- "A piece of data or content is open if **anyone** is **free to use, reuse, and redistribute** it — subject only, at most, to the requirement to attribute and/or share-alike."

OPEN	NOT OPEN
PDB COD,Crystaleye	CCDC, ICSD
RSC/ACS/IUCr CIFs	Elsevier/Wiley/Springer CIFs
Acta Cryst E	Acta Cryst ABCD (default)
CIF dictionaries	

RCUK
Wellcome
ERC
NSF …

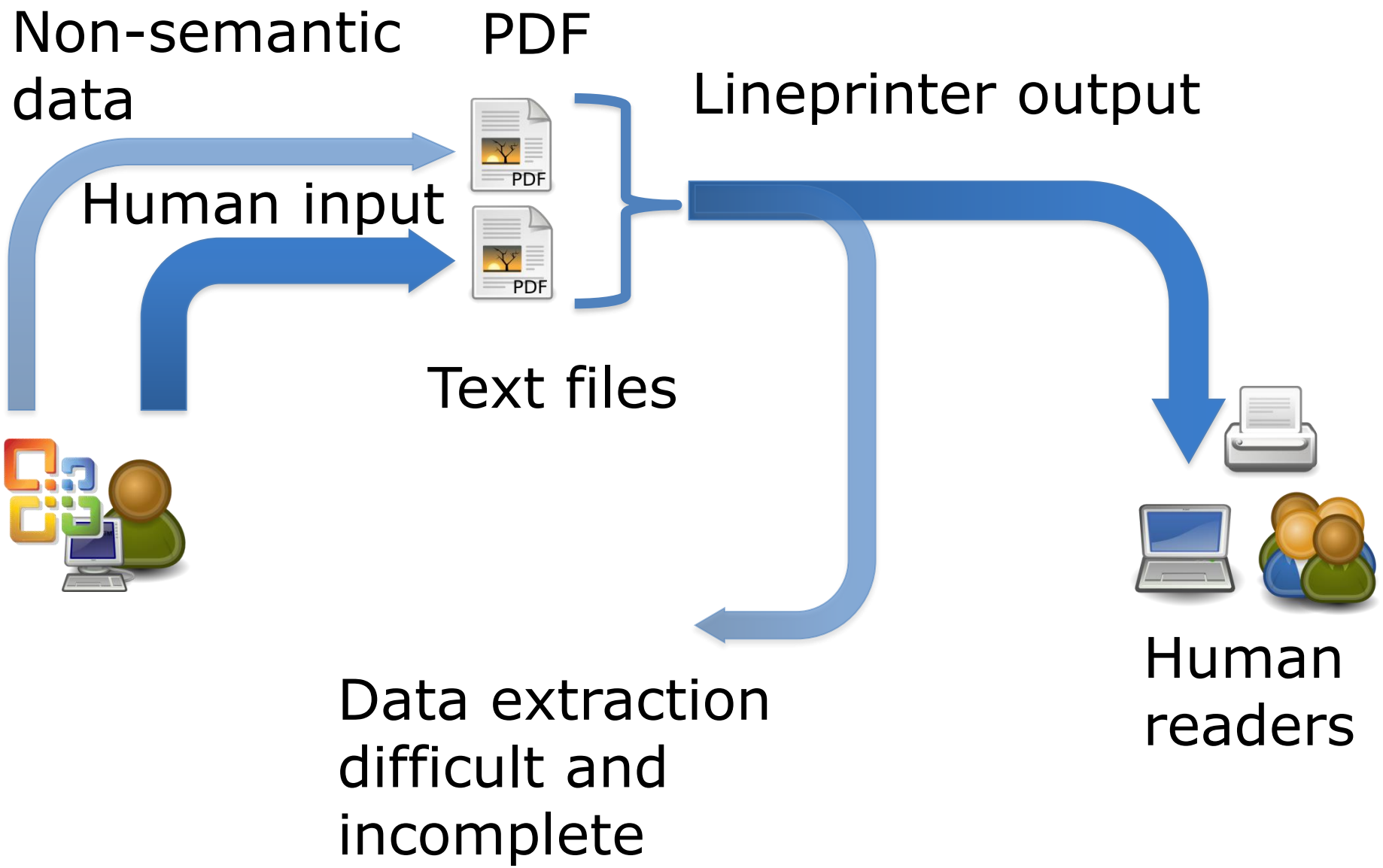
require
fully OPEN

[at Research Data Alliance, we are entering a **new "era of open science",** which will be "good for citizens, good for scientists and good for society".
She explicitly highlighted the transformative potential of open access, open data, open software and open educational resources – mentioning **the EU's policy requiring open access to all publications and data resulting from EU funded research**.

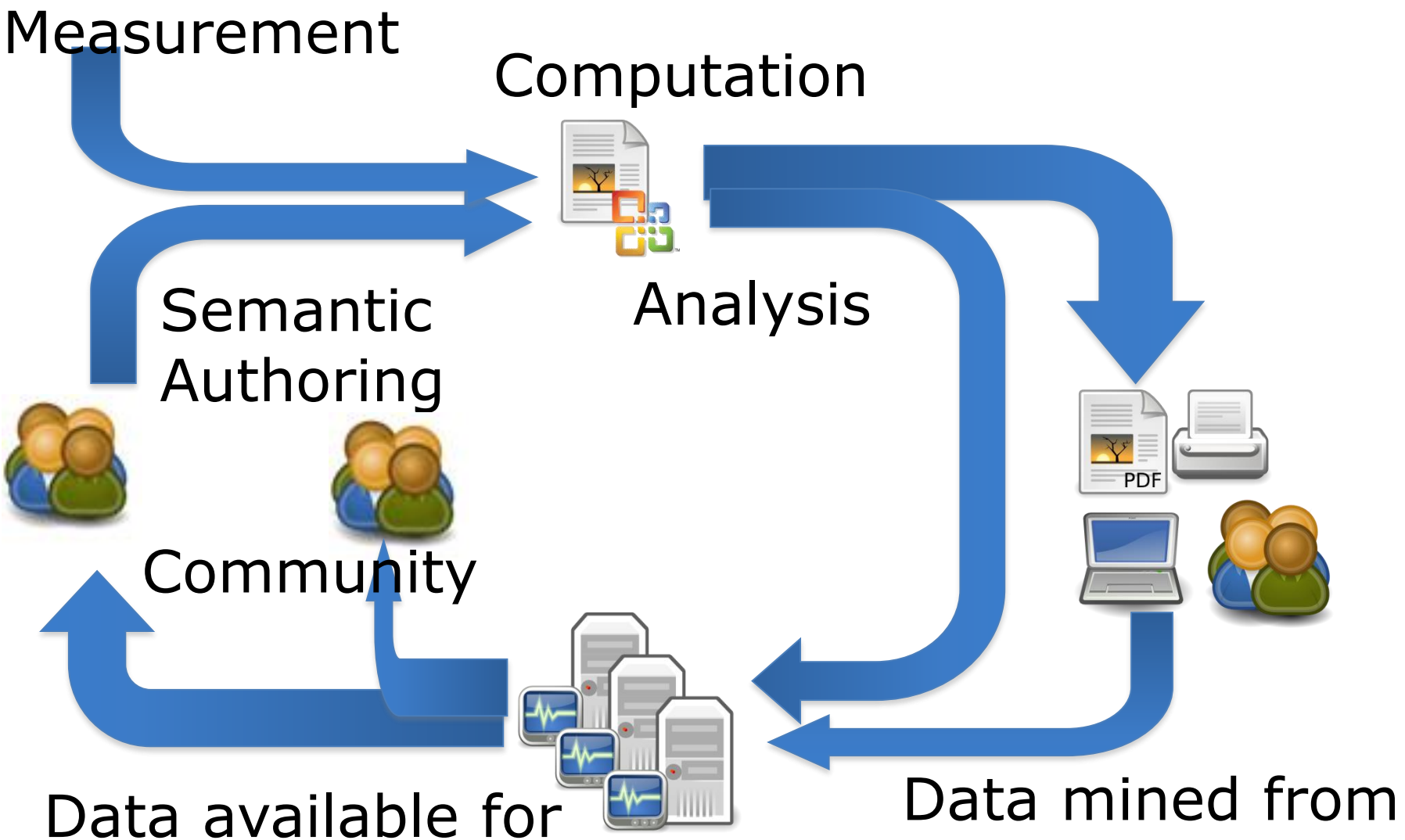
http://blog.okfn.org/2013/03/21/we-are-entering-an-era-of-open-science-says-eu-vp-neelie-kroes/#sthash.3SWDXDE6.dpuf

Current scientific information flow

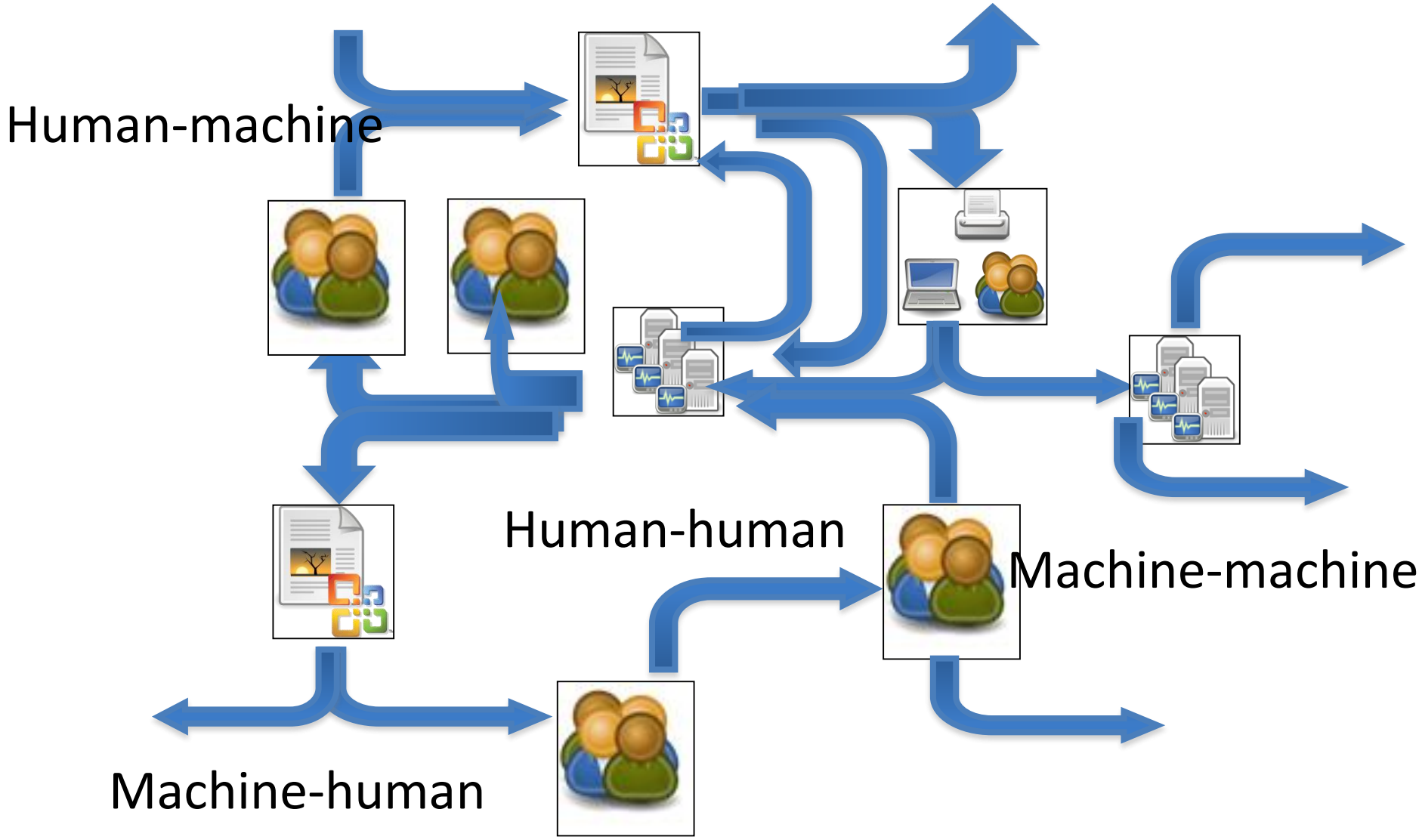
… is broken for data-rich science



Semantic network closes the loop



The network grows autonomously



TimBerners-Lee's Open data

http://5stardata.info

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

CIFDIC
ACS
IUCr

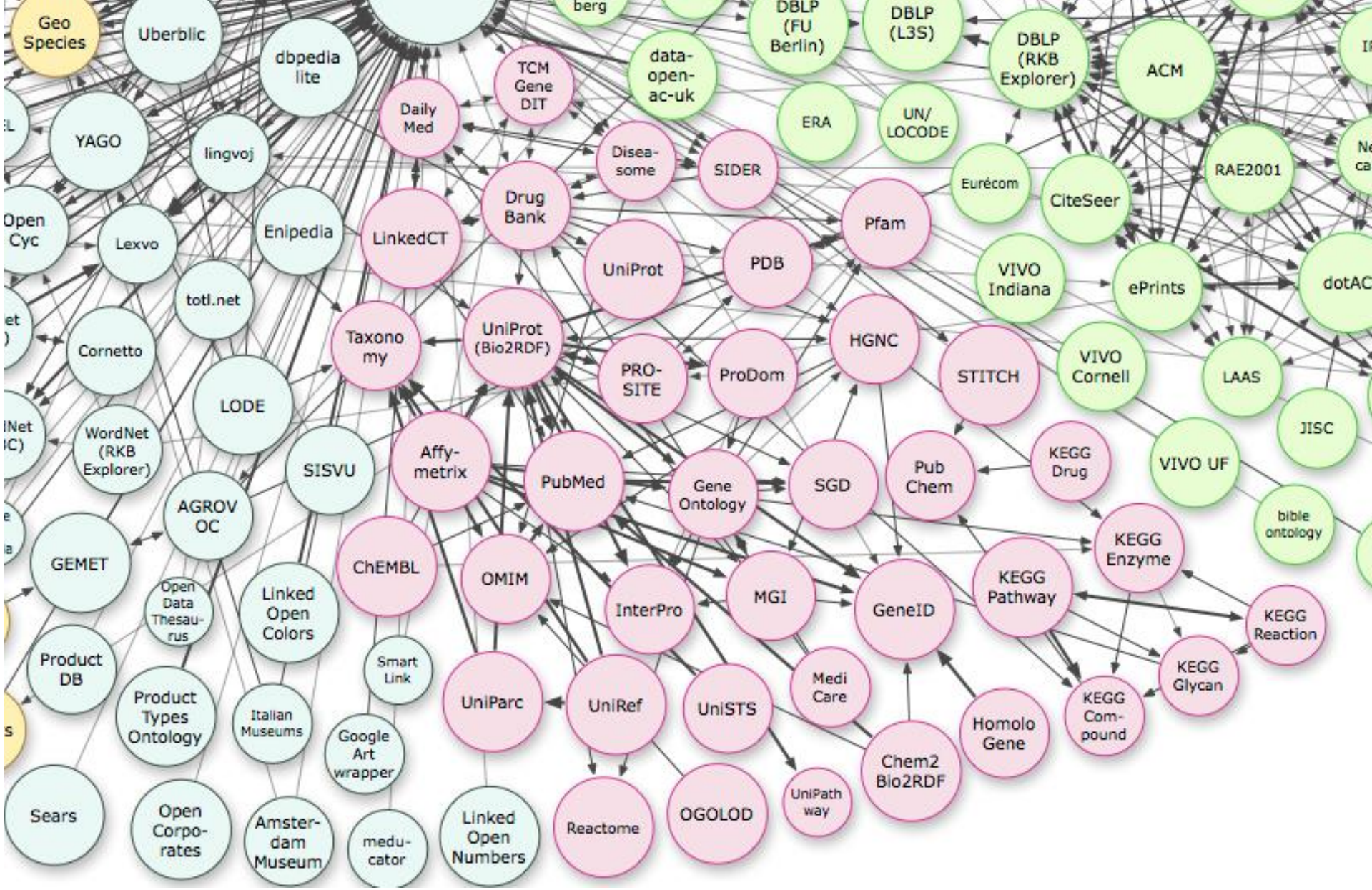
★★ make it available as **structured data** (i.e. NOT PDF)

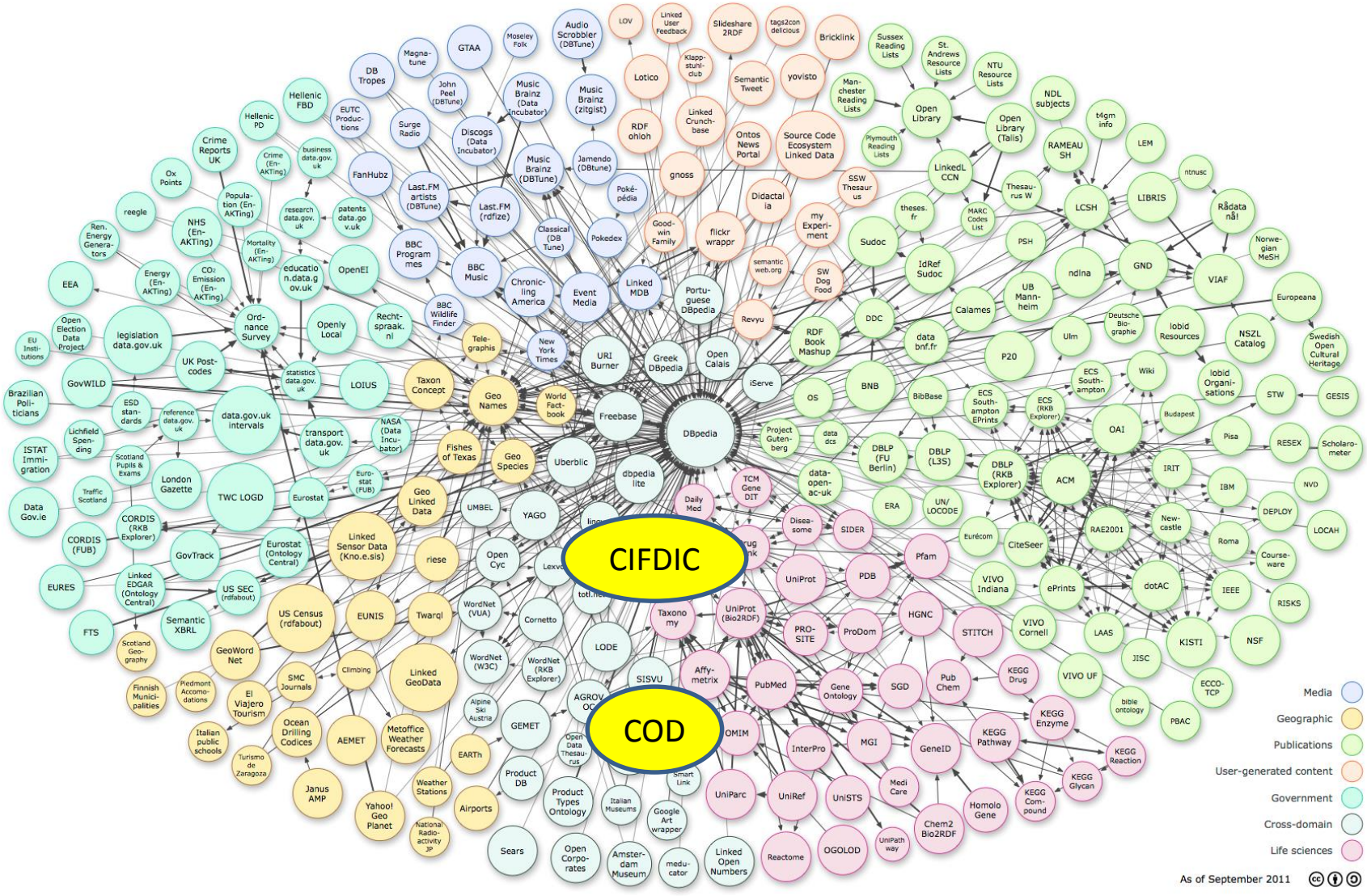
CRYSTALEYE

★★★ use **non-proprietary formats** (e.g., CSV)

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

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


Geo Species
Uberblic
dbpedia lite
YAGO
lingvoj
Open Cyc
Lexvo
Enipedia
totl.net
Cornetto
LODE
WordNet (RKB Explorer)
SISVU
AGROV OC
GEMET
Product DB
Product Types Ontology
Sears
Open Data Thesau-rus
Linked Open Colors
Italian Museums
Open Corpo-rates
Amster-dam Museum
Smart Link
Google Art wrapper
medu-cator
Linked Open Numbers
Daily Med
TCM Gene DIT
data-open-ac-uk
Disea-some
SIDER
Drug Bank
LinkedCT
UniProt
PDB
Pfam
Taxono my
UniProt (Bio2RDF)
PRO-SITE
ProDom
HGNC
STITCH
Affy-metrix
PubMed
Gene Ontology
SGD
Pub Chem
KEGG Drug
ChEMBL
OMIM
InterPro
MGI
GeneID
KEGG Pathway
KEGG Enzyme
KEGG Reaction
KEGG Glycan
KEGG Com-pound
Homolo Gene
Chem2 Bio2RDF
UniPath way
OGOLOD
Reactome
UniParc
UniRef
UniSTS
Medi Care
DBLP (FU Berlin)
DBLP (L3S)
DBLP (RKB Explorer)
ACM
ERA
UN/ LOCODE
Eurécom
CiteSeer
RAE2001
VIVO Indiana
ePrints
VIVO Cornell
LAAS
JISC
VIVO UF
bible ontology


CIFDIC
COD
Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011

Creating semantic content

1. Authoring tools for humans
2. Program output
3. Chemical databases
4. **Content mining and Natural Language Processing (Text) (NLP)**
5. **Community**

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

University of Cambridge › Department of Chemistry › Unilever Centre for Molecular Science Informatics

New: Now works on Atmospheric Chemistry Abstracts (Beta)

ChemicalTagger is an open-source tool that uses OSCAR4 and NLP techniques for tagging and parsing e: the chemistry literature.

To use this demo, select the type of chemistry you would like to analyse (below), enter some chemical t 'Process Text' button

◉ Organic ○ Atmospheric **Typical chemical synthesis**

To a stirred solution of 4-hydroxypiperidine (0.97 g, 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0°C was added 1-(bromomethyl)-4-methoxybenzene (1.93 g, 9.60 mmol) and triethylamine (2.16 g, 21.4 mmol). The reaction mixture was then warmed to room temperature and stirred overnight. After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate (40 mL), washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate. The drying agent was filtered off and the filtrate concentrated under reduced pressure. The residue obtained was purified by flash chromatography (silica gel, 0-5% methanol/methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g, 80%).

Automatic semantic markup of chemistry

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre For Molecular Science Informatics

ChemicalTagger

University of Cambridge › Department of Chemistry › Unilever Centre for Molecular Science Informatics

To a stirred solution of 4-hydroxypiperidine (0.97 g , 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0 °C was added 1-(bromomethyl)-4-methoxybenzene (1.93 , 9.60 mmol) and triethylamine (2.16 g , 21.4 mmol) . The reaction mixture was then warmed to room temperature and stirred overnight . After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate 40 mL) , washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate . The drying agent was filtered off and the filtrate concentrated under reduced pressur . The residue obtained was purified by flash chromatography (silica gel , 0-5 % methanol / methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g , 80) .

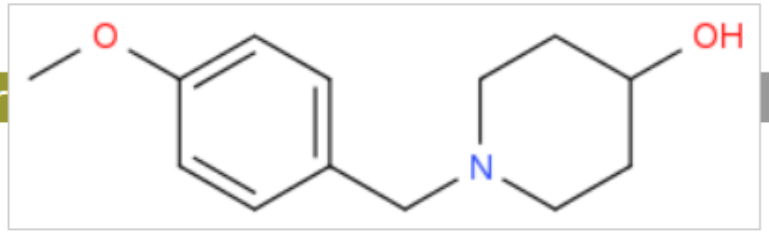
☑ **Actions:**

☑ Filter ☑ Yield ☑ Heat ☑ Wash ☑ Concentr Purify

☐ **Conditions:**

☐ TempPhrase ☐ TimePhrase

☐ **Molecules:**

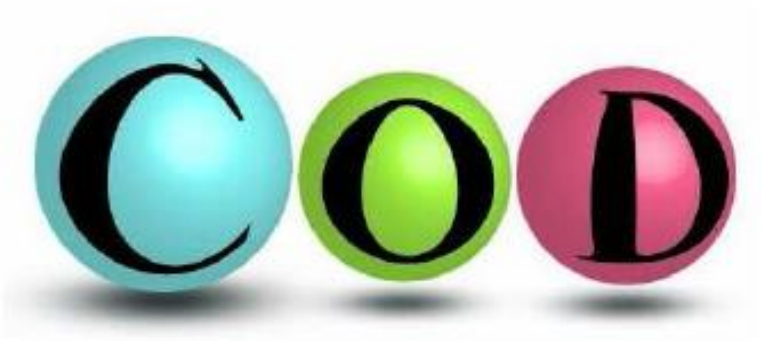


Could be used for analytical, crystallization, etc.

Crystallography Open Database

A grass-root initiative

From Saulius Grazulis



http://www.crystallography.net/

- Total ≈**230 000** records
- **236** registered users
- **31** depositors (deposited at least one structure)
- In year 2012:
  - >**56 000** new structures uploaded (**26 000** more than in 2011)
  - **24** active depositors (who deposited at least one structure in 2012)

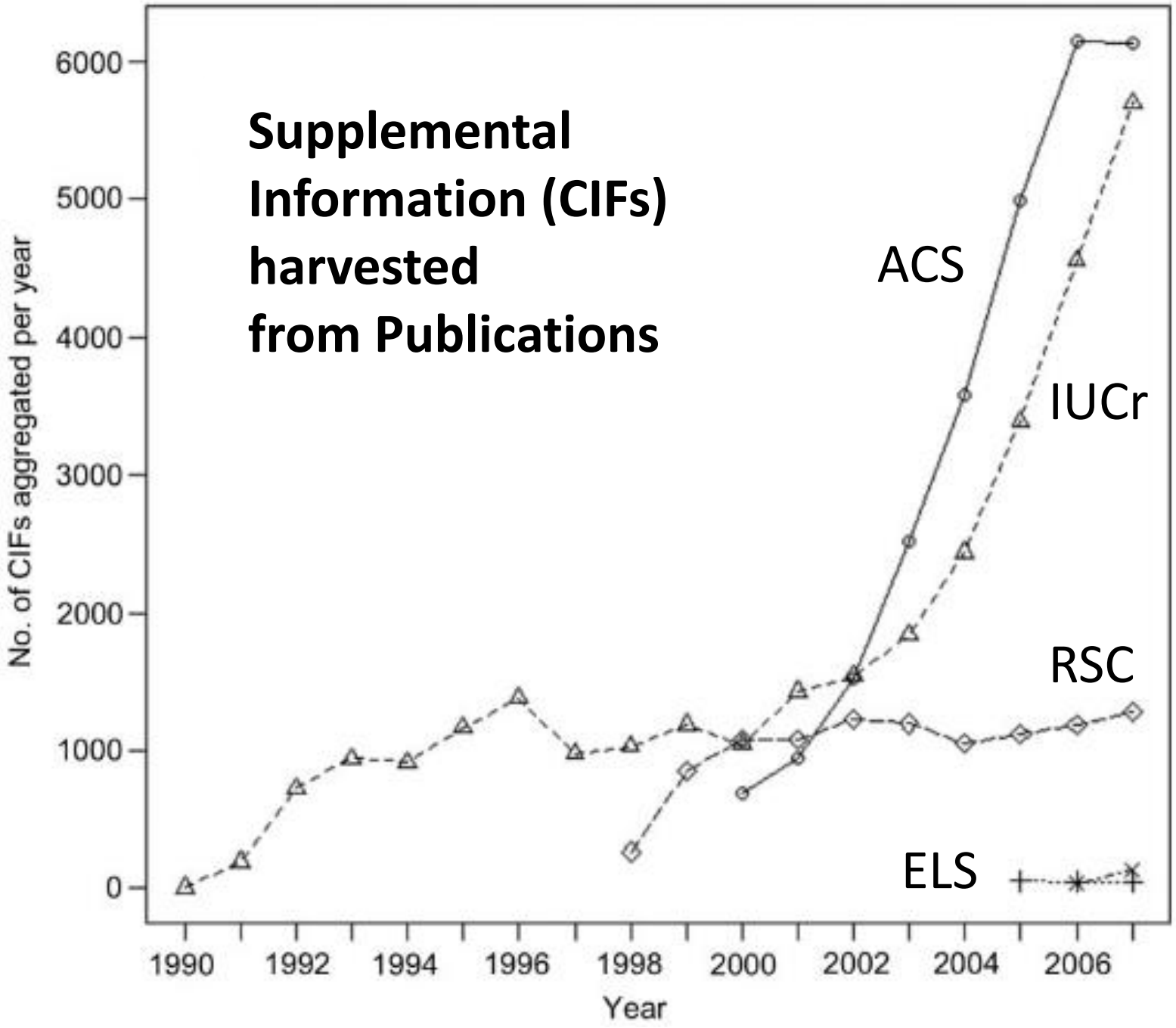
Crystaleye

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

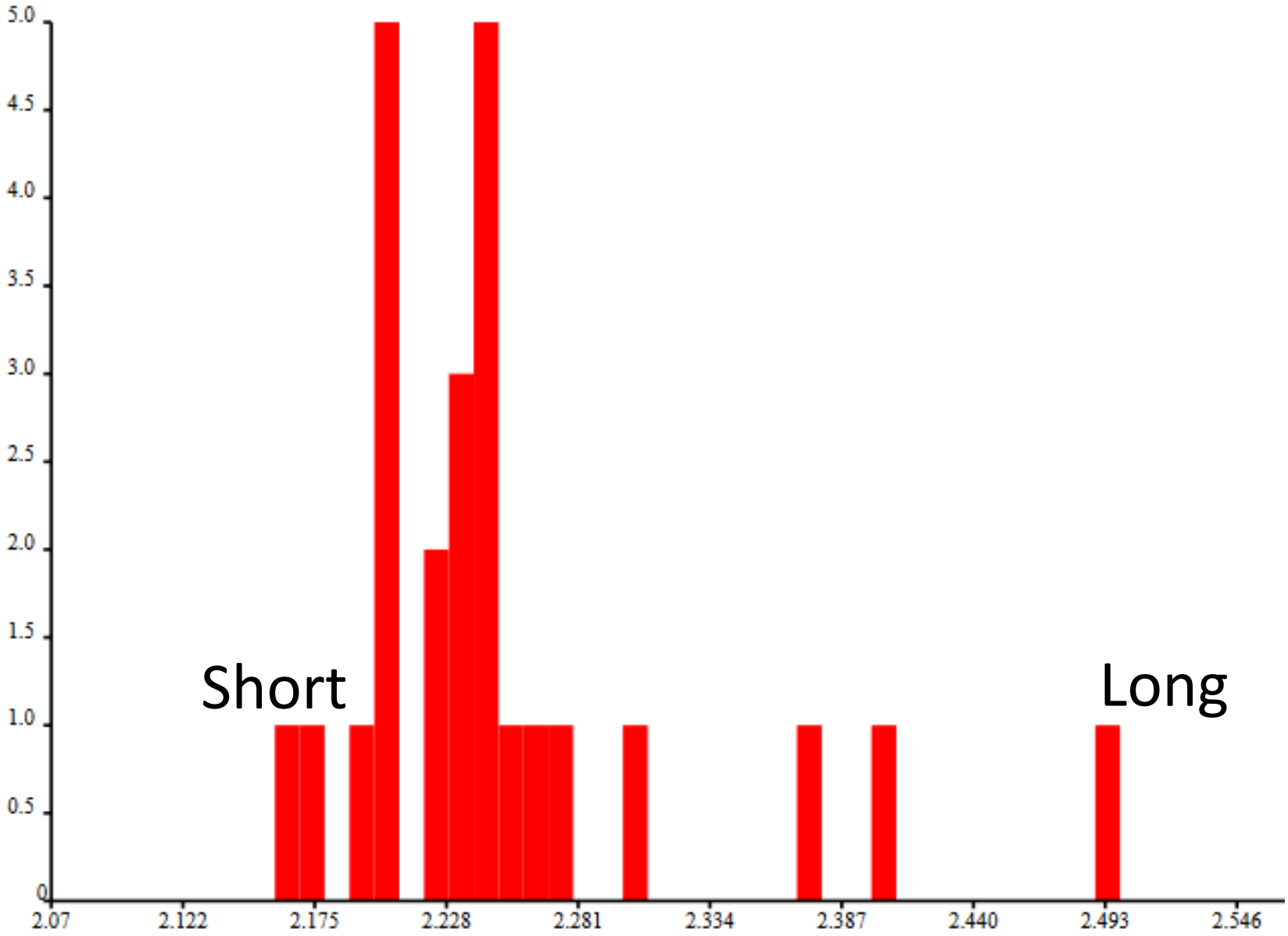
Nick Day, Jim Downing, Sam Adams, N. W. England and Peter Murray-Rust*

J.Appl.Cryst. (2012). 45 , 316–323, doi:10.1107/S0021889812006462

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


Supplemental Information (CIFs) harvested from Publications
ACS
IUCr
RSC
ELS
No. of CIFs aggregated per year
Year
0
1000
2000
3000
4000
5000
6000
1990
1992
1994
1996
1998
2000
2002
2004
2006

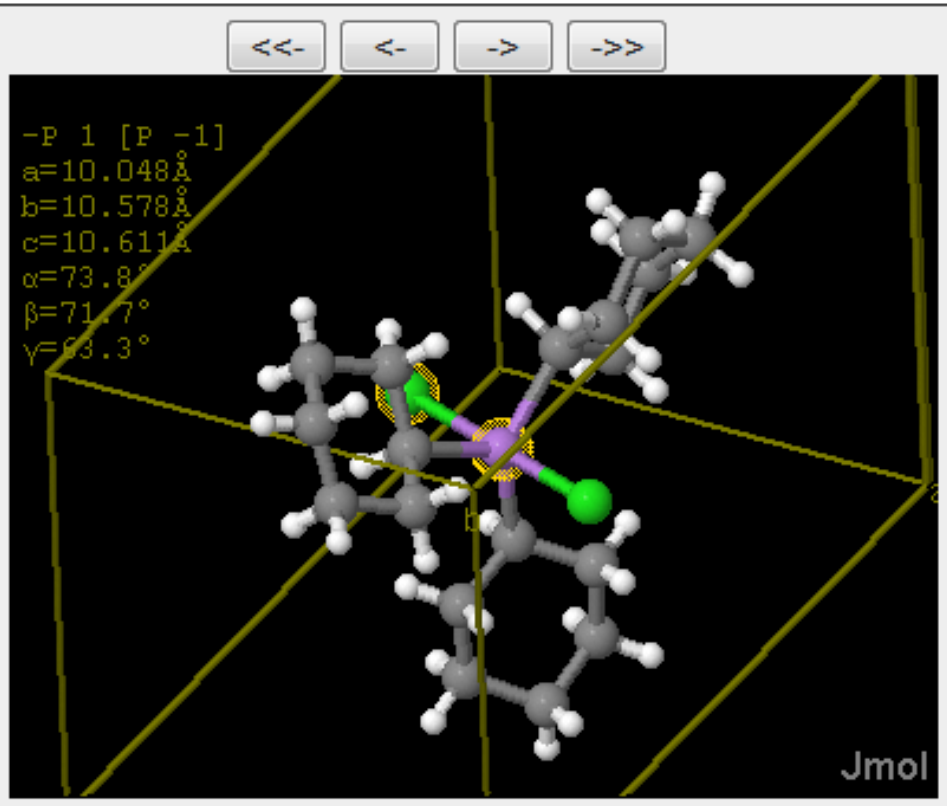
As-Cl Bond lengths

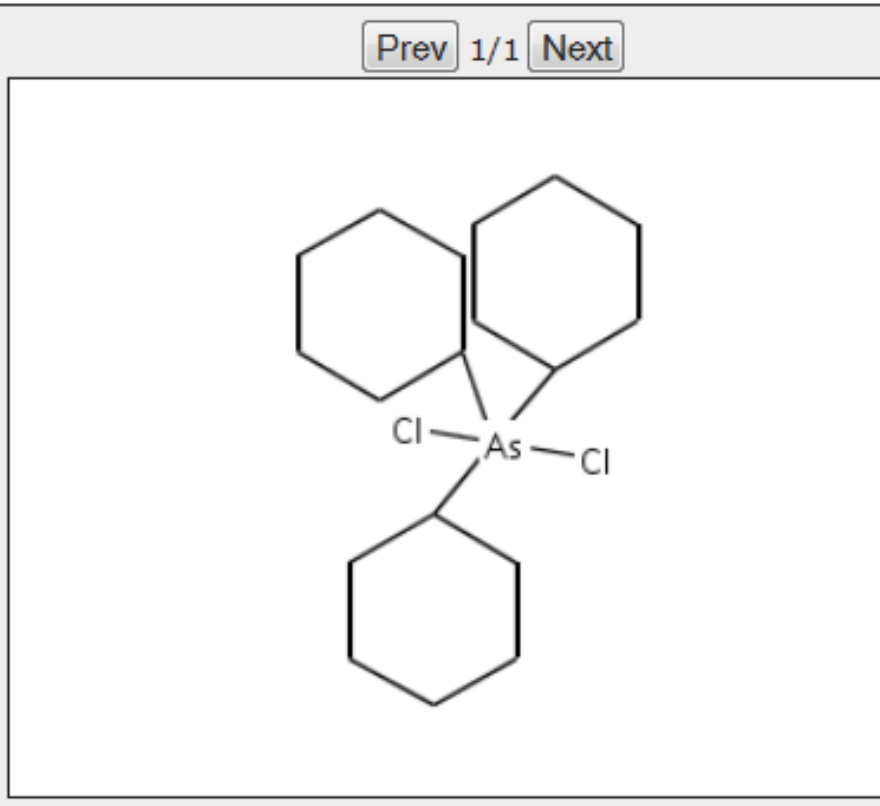


Structures containing As-Cl bonds between 2.49-2.5 Å

Published Formula (clickable)	Article	Summary
$C_{18}H_{33}AsCl_2$	view	view

Long

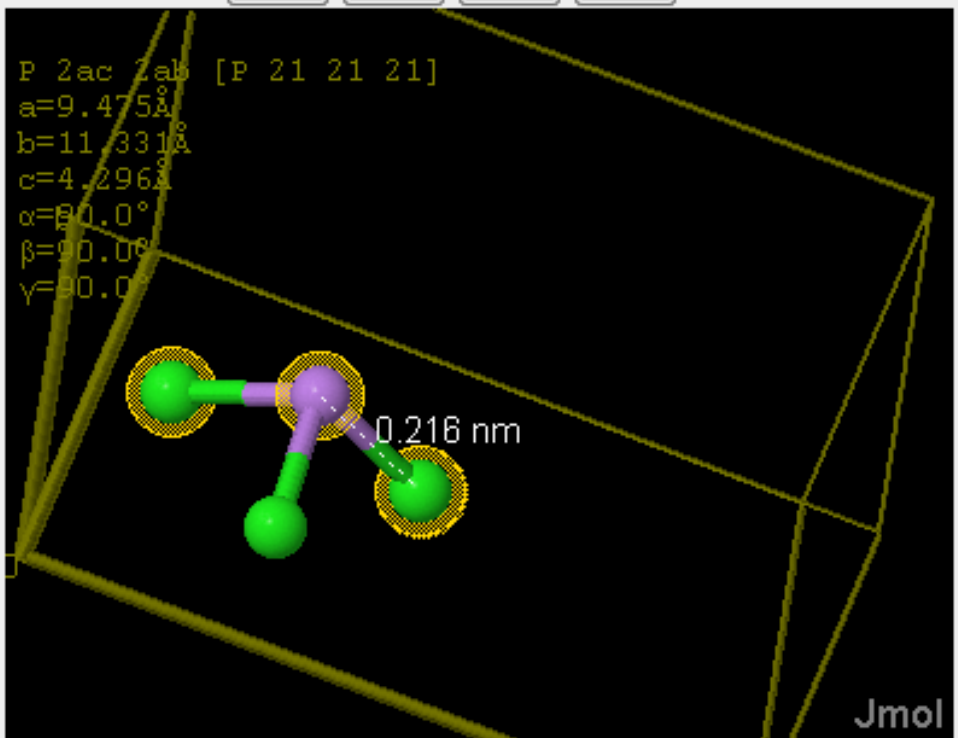


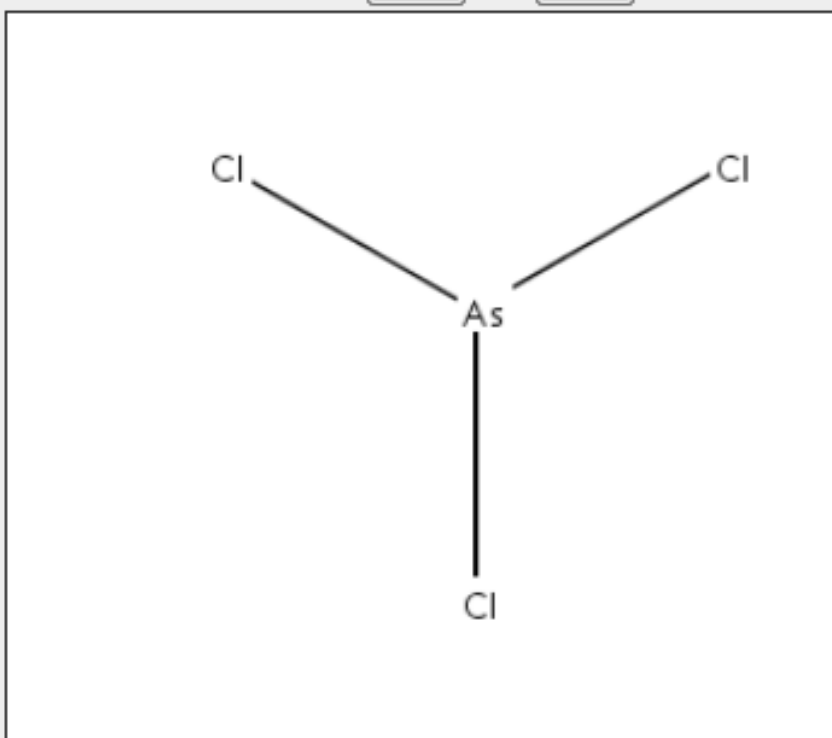


Structures containing As-Cl bonds between 2.16-2.17 Å

Published Formula (clickable)	Article	Summary
$AsCl_3$	view	view
$(C_{16}H_{16}As_2Cl_2S_4)(AsCl_3)_2$	view	view

Short





ROYAL SOCIETY OF CHEMISTRY
ORGANIC AND BIOMOLECULAR CHEMISTRY, 2010, ISSUE 15

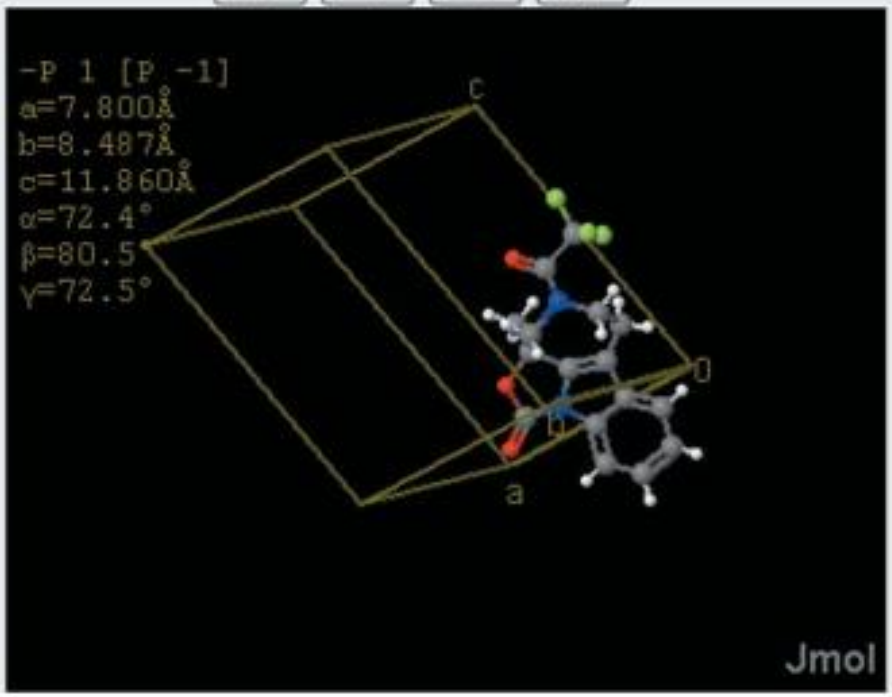
ORGANIC STRUCTURES

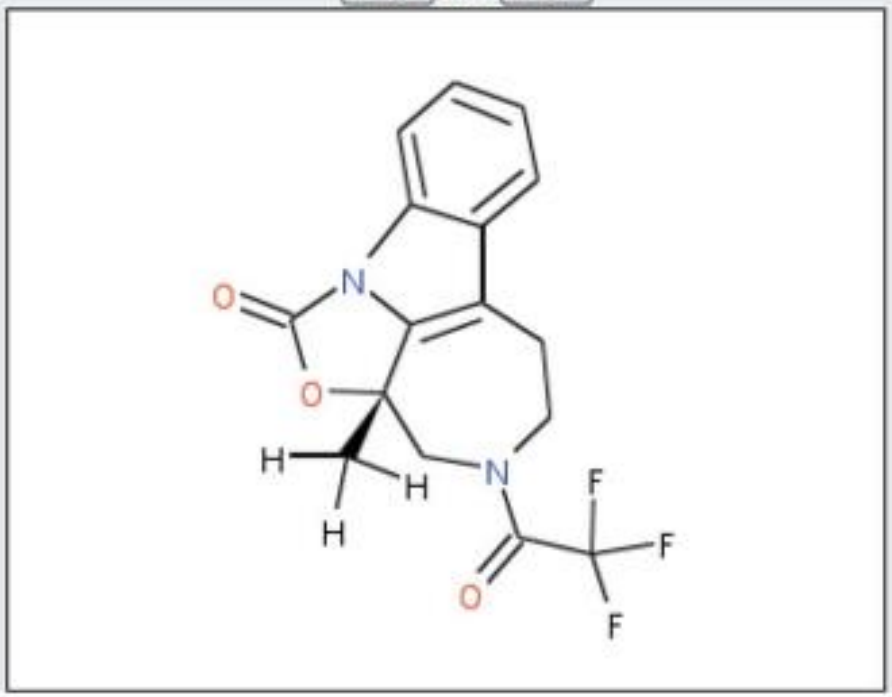
Link to Journal

$C_{16}H_{13}F_3N_2O_3C_{16}H_{13}F_3N_2O_3$	view	view
$C_{15}H_{15}F_3N_2O_4C_{15}H_{15}F_3N_2O_4$	view	view
$C_{24}H_{20}N_4S$	view	view

<<- <- -> ->>

Prev 1/1 Next

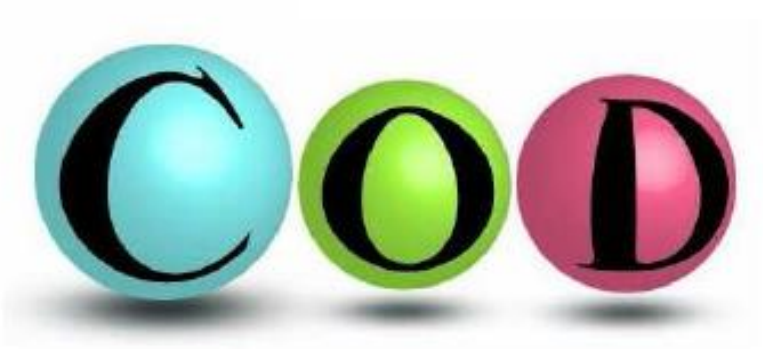




Crystallography Open Database

A grass-root initiative

From Saulius Grazulis



http://www.crystallography.net/

- Total ≈**230 000** records
- **236** registered users
- **31** depositors (deposited at least one structure)
- In year 2012:
  - >**56 000** new structures uploaded (**26 000** more than in 2011)
  - **24** active depositors (who deposited at least one structure in 2012)

COD Letter to Editors 2012

[We] have become aware of **growing concerns** regarding the publication, preservation and quality maintenance **of crystallographic data**. …However, we believe that **completely open deposition** of data and multiple checks can ensure the quality and wide availability of scientific data

[Please] recommend to your authors that, **they also deposit their supplementary crystallographic data into the COD** when they submit scientific papers to your journals.

Being **open by its design**, the COD enables the creation of multiple mirrors and backup copies. It provides, thus**, archival storage** of scientific data with adequate reliability. … services for **reviewers and editors** to facilitate the peer-review. …**since our database follows the Open Access model**, all material deposited into the COD is **available to other databases**. The COD team actually encourages the use of our data collection for any possible scientific or industrial application by putting the **database into the public domain**

Recommendations for Open Crystallography

- Require Open Crystal Data for all publications
- Deposition of Open Data in COD
- Integrate CIF dictionaries as RDF into Linked Open Data
- Integrate COD into Linked Open Data Cloud
- CCDC/ICSD to publish RAW author CIFs Openly