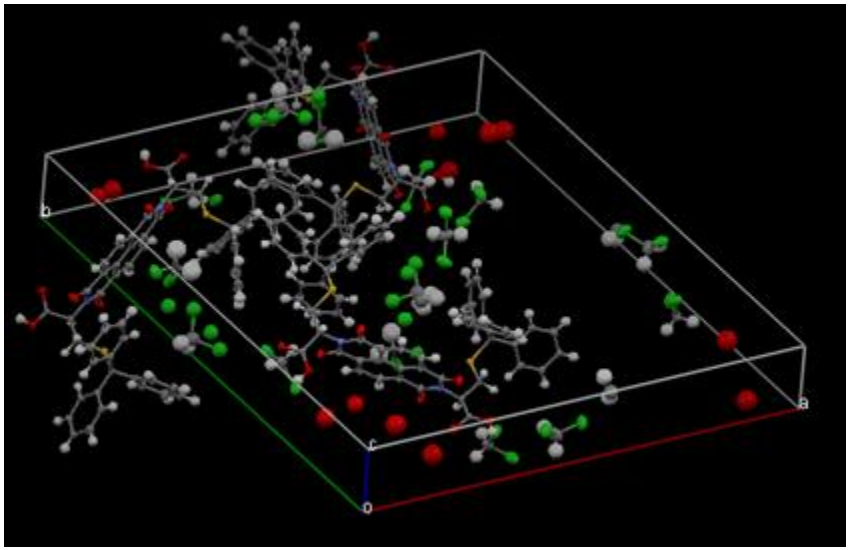
JISC XYZ Project



Managing Research
Data Workshop

Birmingham
28/29 March 2011

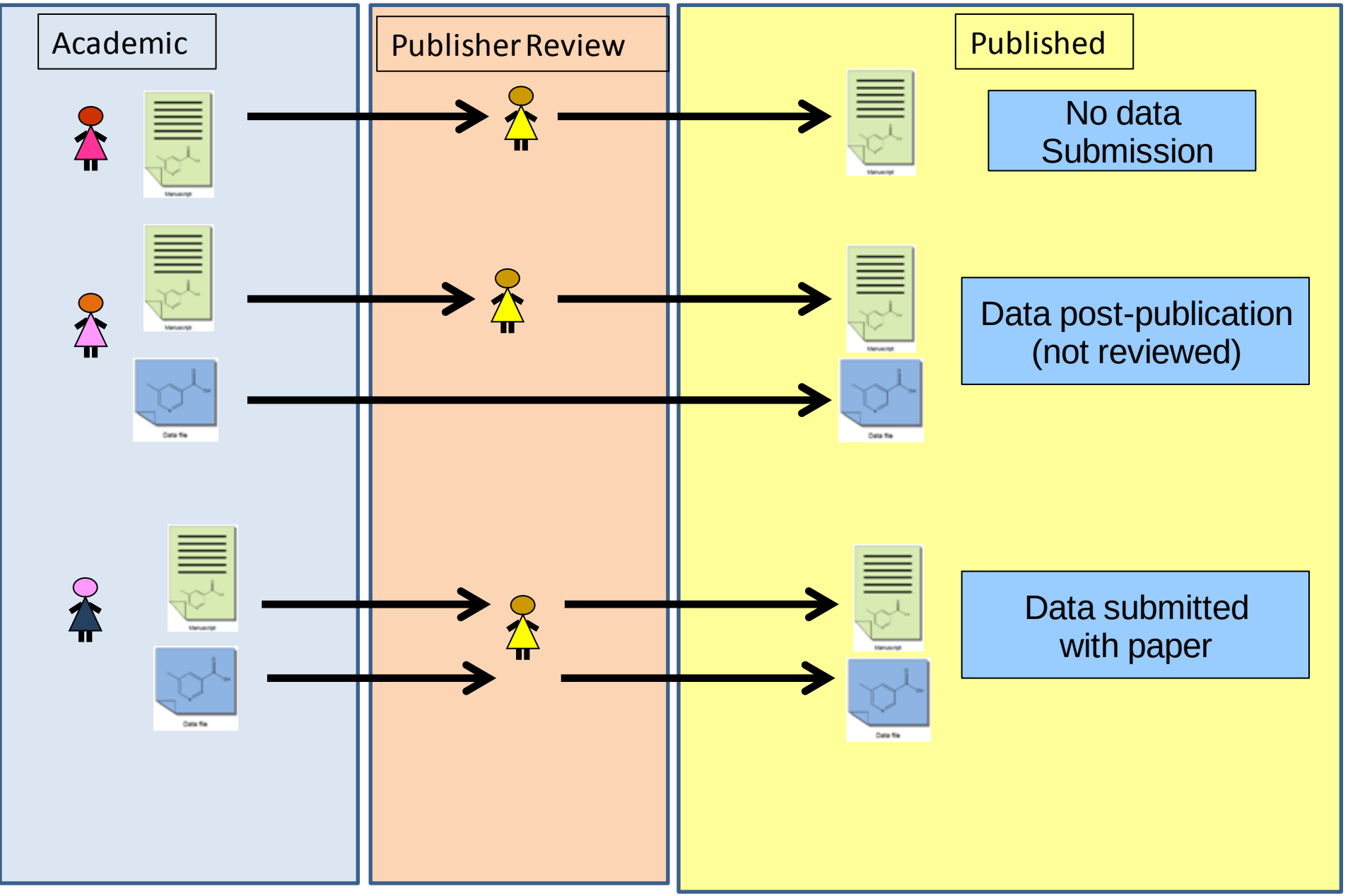
Principal Investigator: ***Peter Murray-Rust***
Project Team: ***Nick England, Brian Brooks***
Unilever Centre, Department of Chemistry, University of Cambridge

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre For Molecular Science Informatics

Publishing scientific data

- Challenge: How can scientists be encouraged to provide data in support of their papers?
- Academic papers:
  - Publication record is important to academics
  - Papers rarely have supporting data files
  - Asking for data post-publication not optimal
- **XYZ project** – ask for data up-front when a paper is submitted
  - Provision of data is a condition of acceptance of the paper



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre For Molecular Science Informatics

Acta Cryst E

- IUCr = International Union of Crystallography
  - Strong supporters of Open Data
- Acta Cryst E is primarily for publishing crystallographic **data**
- XYZ project working with IUCr
- Building a data journal (i.e. a publication which contains data)

Acta Cryst E

- Example of Best Practise
- Data submitted with paper
- Automatic validation of data
- Data and validation report available to reviewers.
- Data available for download when the paper is published

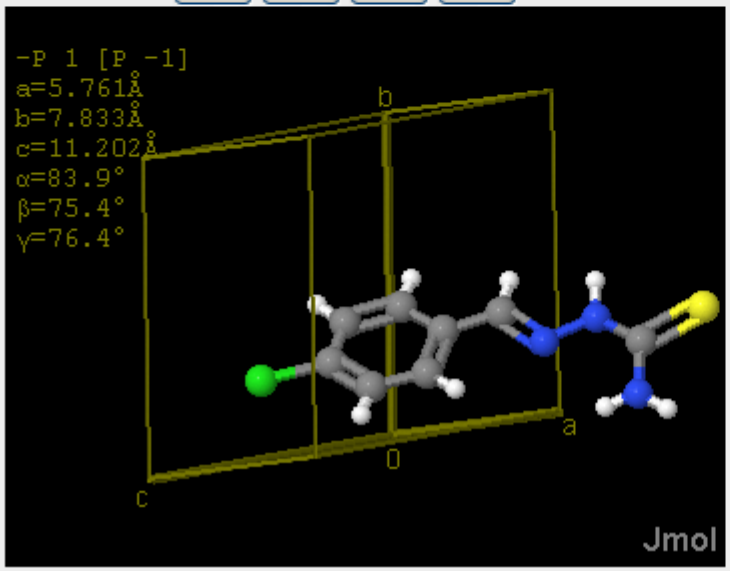
Advantages of Data?

Acta Crystallographica
Section E: Structure Reports, 2011, issue 03-00

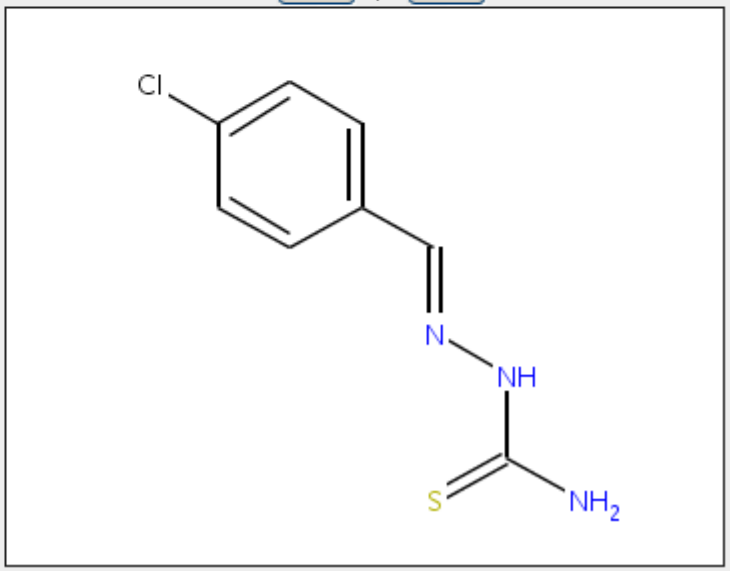
Organic Structures

$C_8H_8ClN_3S$	view	view
$(C_{18}H_{21}NO_3)(H_2O)_{2.33}$	view	view
$C_{14}H_{11}FN_2O_2$	view	view
$C_{17}H_{18}N_2O_2$	view	view
$C_{29}H_{25}N_3O_3$	view	view
$C_{32}H_{30}Cl_4N_4$ ((DP))	view	view
$C_{18}H_{17}N_3$	view	view
$C_{28}H_{37}NO_3$	view	view

<<- <- -> ->>



Prev 1/1 Next



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre For Molecular Science Informatics

What format should data be stored in?

- PDF? XML? RDF? XHTML? Other…?
- PDF:
  - Commonly used
  - Used for long-term archiving
  - Great for printing, reading; not good for data retrieval
  - 100% people-oriented - Horrible for machines to read, text mine
- XML:
  - Content only
  - Not easy to author
  - Not compelling for users; good for machines
- RDF:
  - Semantic format – Web3 – Resource Description Framework
  - Good for machines; not good for users

Formats for data (contd)

- HTML:
  - Created in early ‘90’s
  - Combines content and formatting
  - Pervasive; good for humans, good for browsers
  - Not ideal for data storage & use
- XHTML:
  - Extensible HTML – next-generation HTML
  - Store data within HTML as XML or other format
  - Good for humans; good for computers to use content
- RDFa:
  - Resource Description Framework-in-attributes
  - Inclusion of data into XHTML in RDF format
  - Deliver data in RDF format along with the HTML page

Scholarly-HTML (ScHTML)

- Started by Peter Sefton
  - Beyond-The-PDF meeting, California, January 2011
  - Hackfest , Cambridge, March 2011
- Goal for ScHTML is to make things:
  - Improve the scholarly document
  - Without putting too much extra burden on the author
  - Metadata, using a linked data approach.

http://scholarlyhtml.org/

Desirable properties of ScHTML

- Easy to create/store
- Easily accessible
- Easy to extract data
- Store any dataset
- Maintain the digital format of the data (i.e. allow the “filetype” to be conserved)
- Add semantics to storage of data & metadata
- Usable by variety of applications/viewers
- Packaging – able to encapsulate a variety of display and/or data objects in a way that facilitates distribution/ communication of those objects

Principles for ScHTML

- Data is stored in HTML pages
  - Can be part of a more general page, or the page could be purely data
- Data is stored as RDFa
- Use redirection technique to allow different applications to process the same content

ScHTML Example

Beyond the PDF

... built using **ReDBox**... on **The Fascinator** platform.

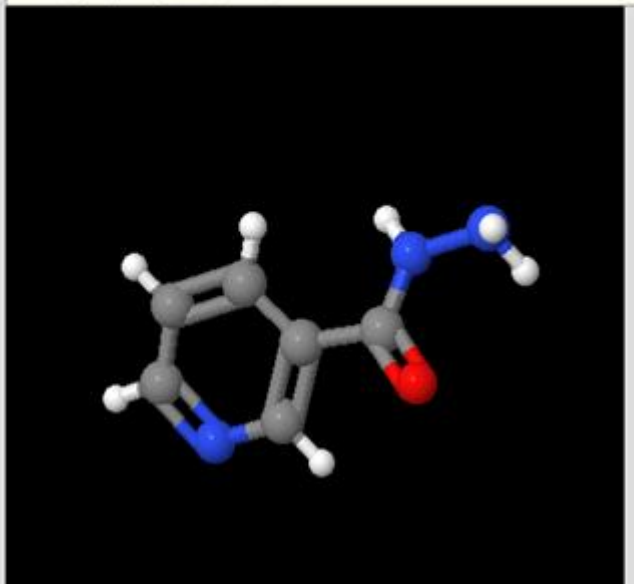
Home **Browse** **Views** **About** **Login** View: Everything

1 mmol of the title compound (purchased from Sigma-Aldrich at 97% purity) was dissolved in a mixture benzene/ethanol (8:1, 50 ml) and refluxed for 1 h. After cooling the solution to ambient temperature, a colorless precipitate was formed, which was collected by filtration and washed with benzene/ethanol (8:1). Crystals suitable for single-crystal X-ray diffraction were grown from a benzene solution, by slow evaporation of the solvent.

PDF version

Comment by: *Peter Sefton* **10 days ago**

fsdjf sdkjf sdkfj sdklfjds flkjds fsdf



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre For Molecular Science Informatics

Demo

- Pete Sefton's demonstration of the potential of ScHTML
- The demonstration:
  - Uses a OpenOffice word document
  - User simply adds a link to the datafile
  - Place the file and datafile in a dropbox folder
  - That's it!

XYZ Project

- Build an Acta Cryst E data journal
- In ScHTML
- Data in semantic form, online
- Convert 250K existing crystal structures from Crystaleye
  - Plus new crystal structures

References

- Peter Murray-Rust blog
- http://blogs.ch.cam.ac.uk/pmr
- Background to ScHTML
- http://tinyurl.com/62zucg8
- Peter Sefton blog
  - http://ptsefton.com/
- ScholarlyHTML
  - http://scholarlyhtml.org
- CrystalEye
- http://wwmm.ch.cam.ac.uk/crystaleye