CrystalEye

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

Nick Day
ned24@cam.ac.uk

Unilever Centre for Molecular Informatics, University of Cambridge, UK.
August 23th, 2008

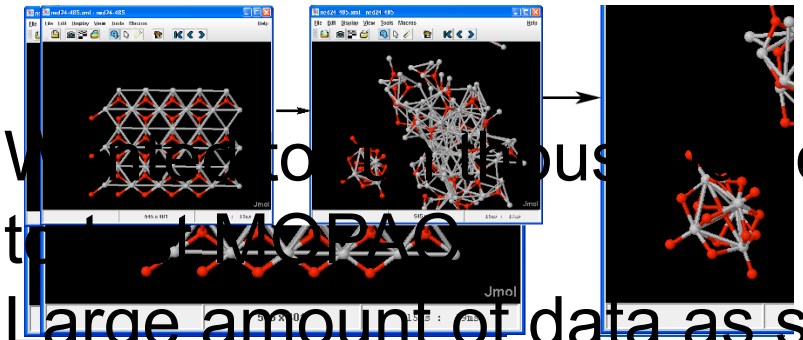
UNIVERSITY OF CAMBRIDGE

Unilever Centre for Molecular Informatics

Crystallographic Data as a Research Resource

Can MOPAC predict crystal structures?

Experimental Calculated



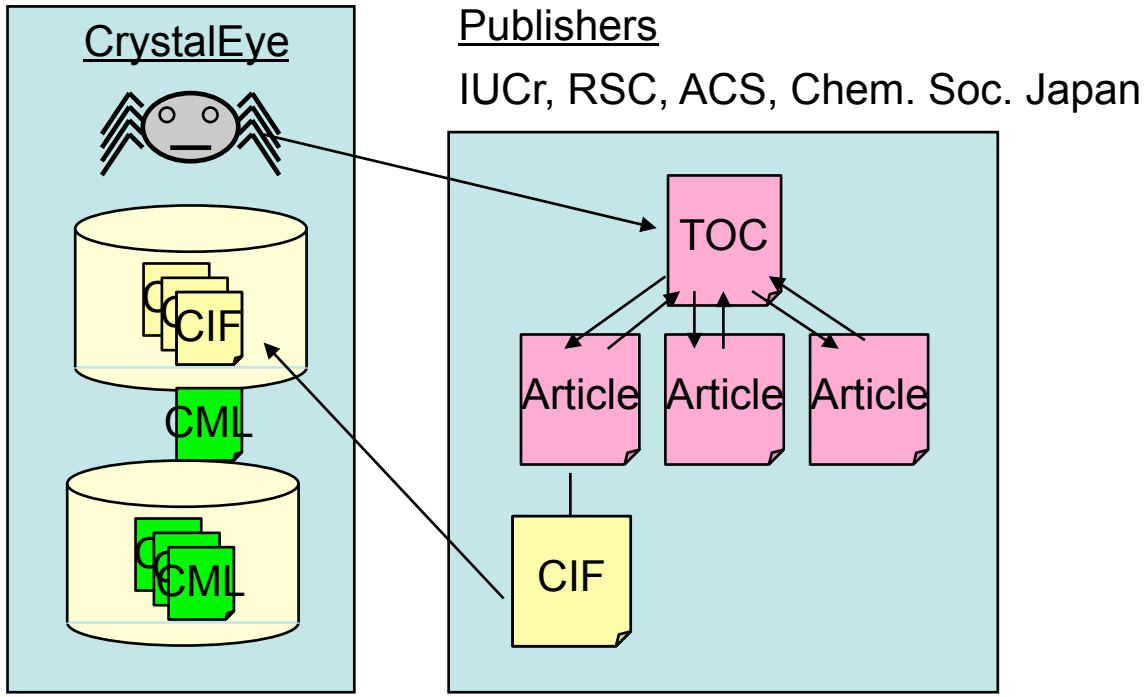
- W[illegible] to [illegible]ous [illegible] of calculations t[illegible] MOPAC.
- Large amount of data as supplementary information in CIF form, but…
    - No easy access to them.
    - No connection tables within.

CrystalEye – a semantic knowledgebase

- A completely automated knowledgebase for e-Science that would:

Aggregation – Today

- Web spider checks publisher's sites every day.
- Parses HTML pages, extracts links and follows them.
- Currently over 88,000 validated CIFs (>120,000 structures).



- Also aggregate from the Crystallography Open Database.

Aggregation – Current Developments

- Departmental Repository for Crystallography, **C3DeR**

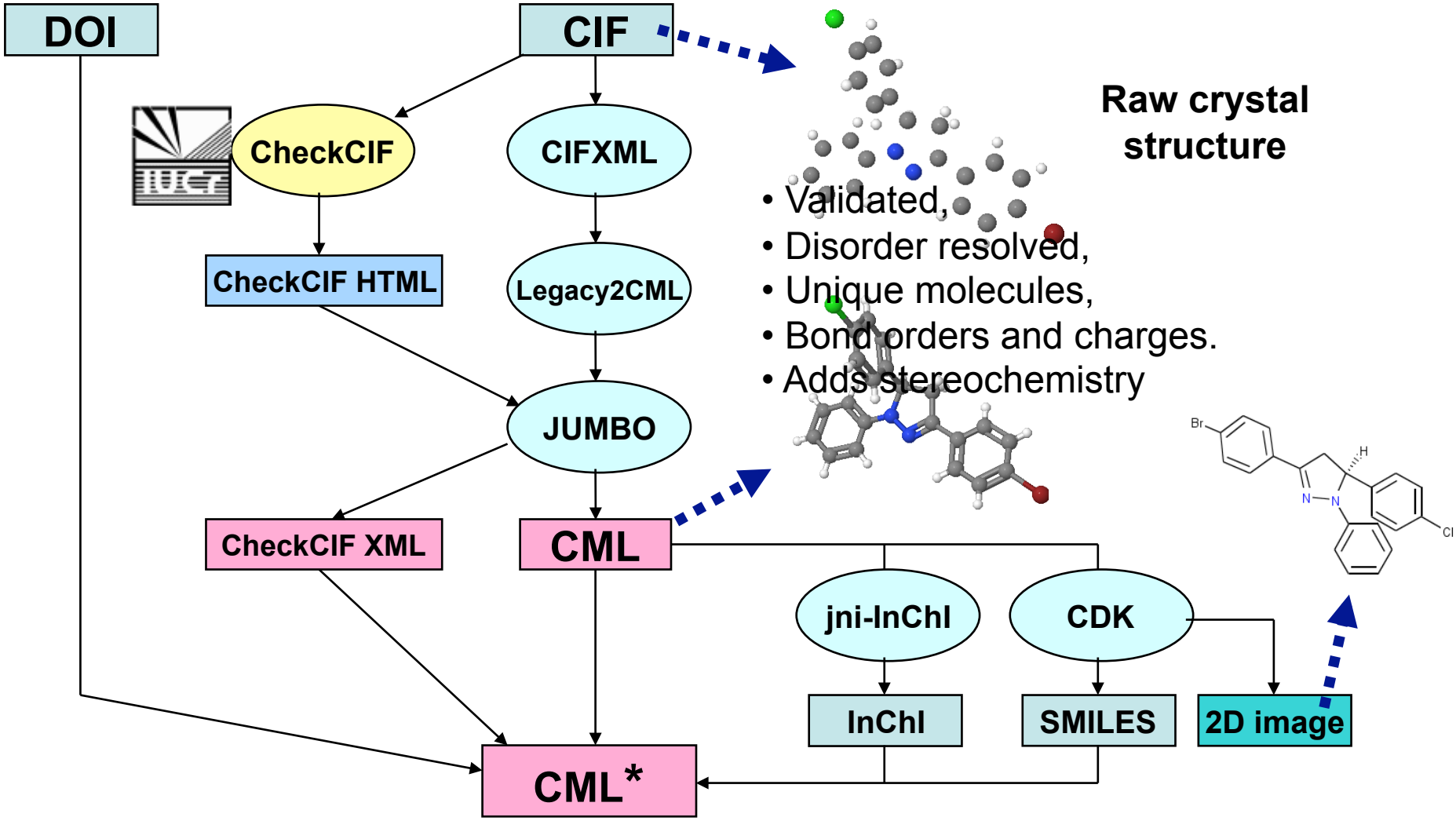
Number of Structures:		269		P	Published			
Note: Dates only start from 15/03/01				D	Unpublished, can be released to department			
				G	Unpublished, can be released to group			
Key:		Data on CIF		E	Structure Embargoed			
		Calculated by computer						
		Information from literature						
		Information supplied by group						
Department ID	DOI	Date	Status	Space Group	Sum Formula	Group Leader	Chemist	Crystallographer
wj0440	10.1107/S1600536805008779	20/12/2004	P	P2(1)/c	C10 H12 N2 O7 S	Bill Jones	A. V. Trask	J. Davies
wj0302	10.1107/S1600536804004453	21/05/2003	P	P2(1)/n	C20 H18 Cl N O4	Bill Jones	A. V. Trask	J. Davies
wj0002	10.1107/S1600536803004446	15/03/2001	P	Pca2(1)	C13 H11 N O4	Bill Jones	N. Shan	J. Davies
wj9904	10.1023/A:1015067422886	15/03/2001	P	P b c a	C12 H20 N2 Ni O4 S6	Bill Jones	A. Bond	A. Bond
wj9905	10.1023/A:1015067422886	15/03/2001	P	P b c a	C12 H20 Co N2 O4 S6	Bill Jones	A. Bond	A. Bond

- Part of the eCrystals federation.

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

Processing – Workflow



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

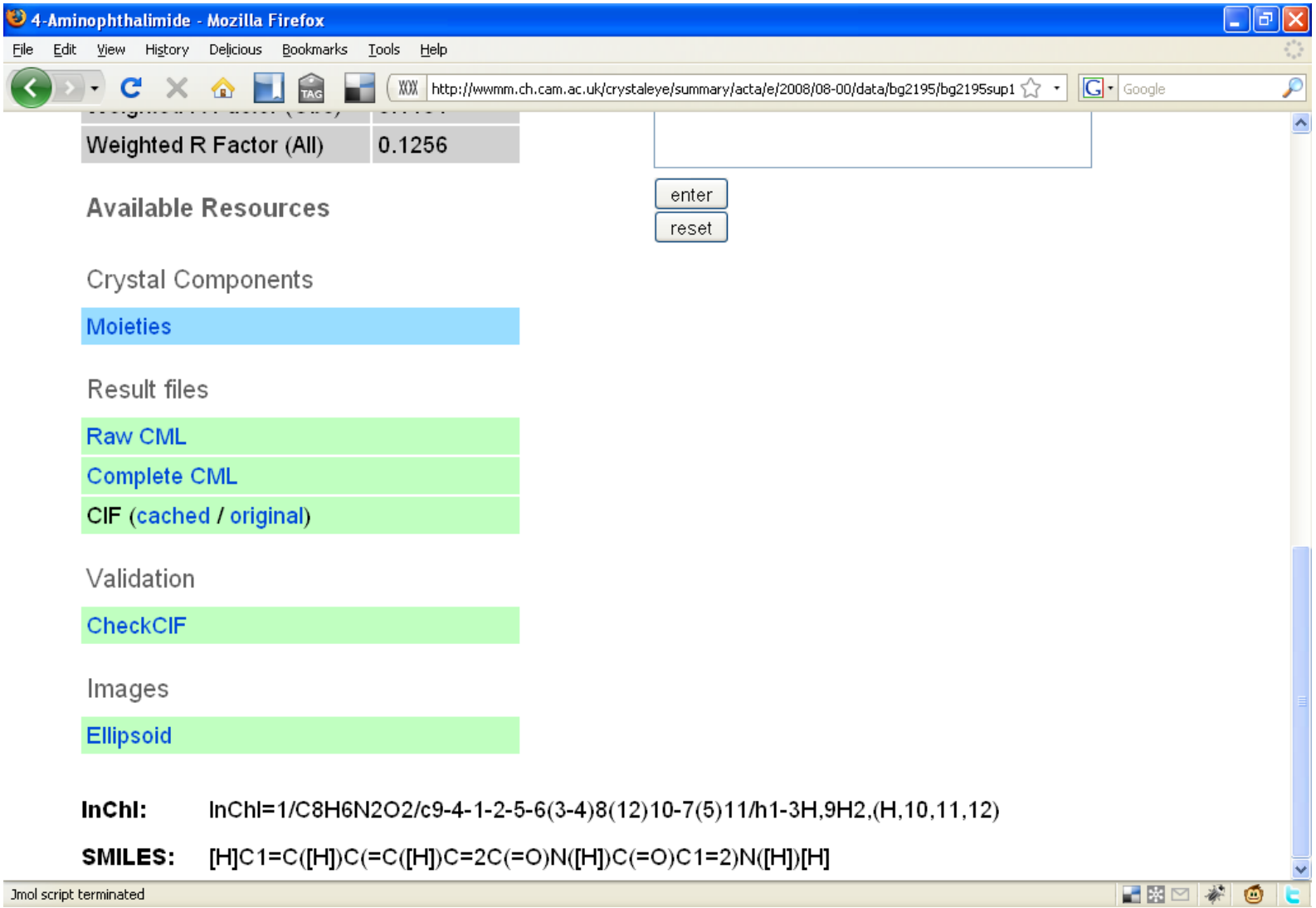
Dissemination

1. Browse
2. Search
3. RSS feeds
4. Harvesting
5. Example services

Dissemination

1. **Browse**
2. Search
3. RSS feeds
4. Harvesting
5. Example services

Dissemination 1 – Browse



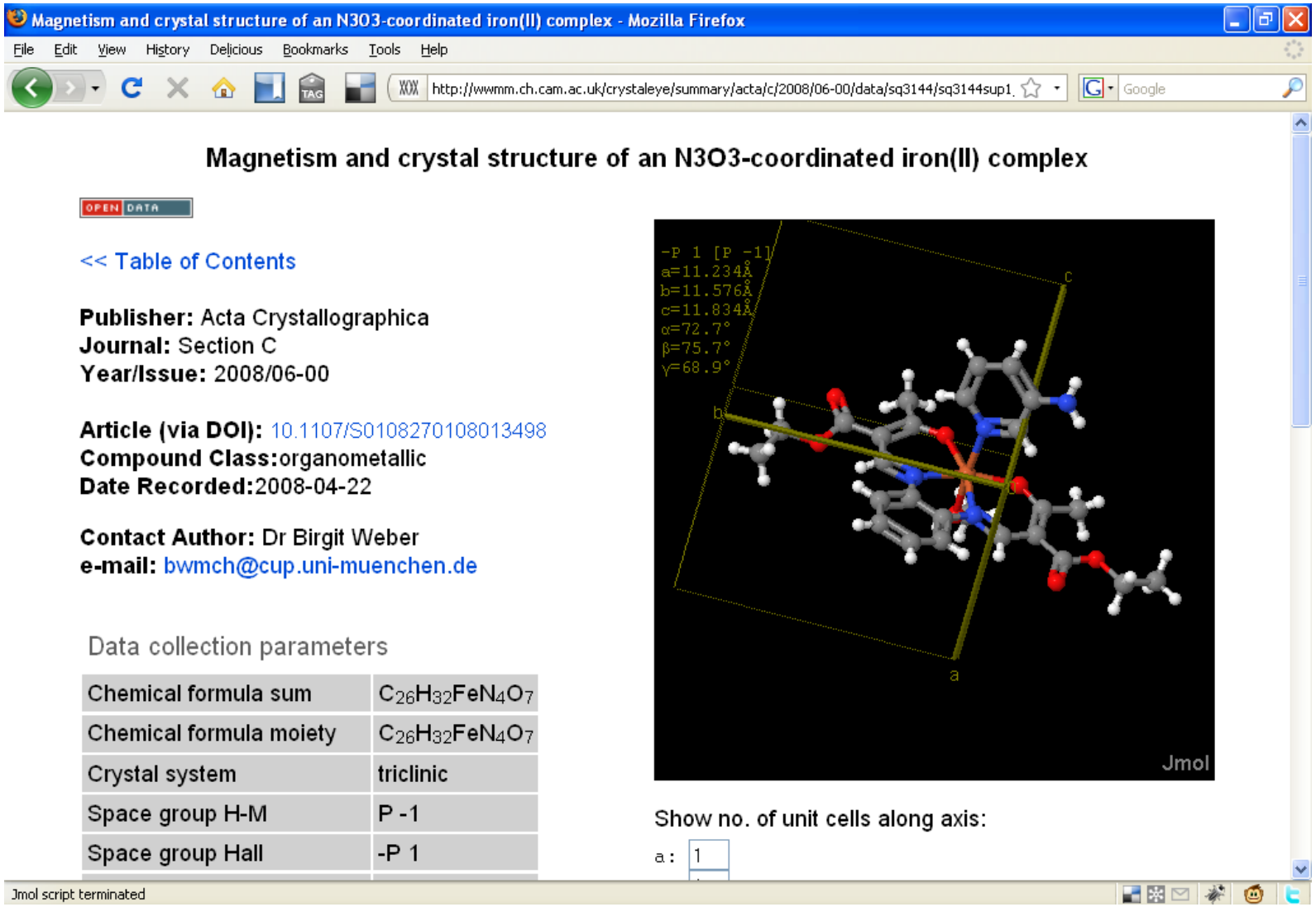
UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

Dissemination

1. Browse
2. **Search**
3. RSS feeds
4. Harvesting
5. Example services

Dissemination 2 – Search



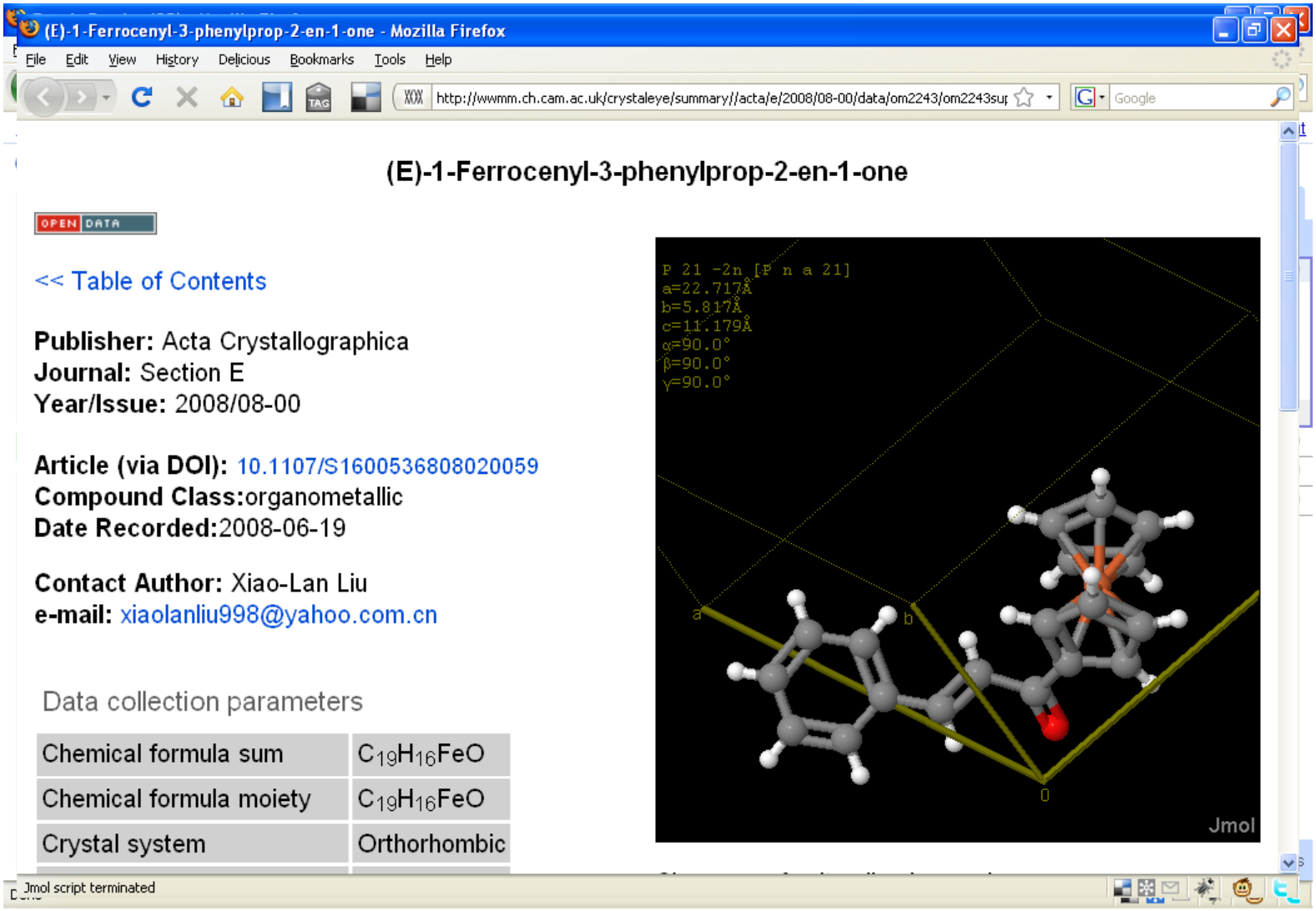
UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

Dissemination

1. Browse
2. Search
3. **RSS feeds**
4. Harvesting
5. Example services

Dissemination 3 – RSS Feeds



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

Dissemination

1. Browse
2. Search
3. RSS feeds
4. **Harvesting**
5. Example services

Dissemination 4 – Harvesting

- Users wanted to harvest the entire CML collection using RSS feeds.
- However, no standard way for harvesters to discover or recover from missed items using standard RSS.
- Atom Archive feeds (RFC5005) extend Atom to fix these problems.

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

Reuse of Harvested CrystalEye data

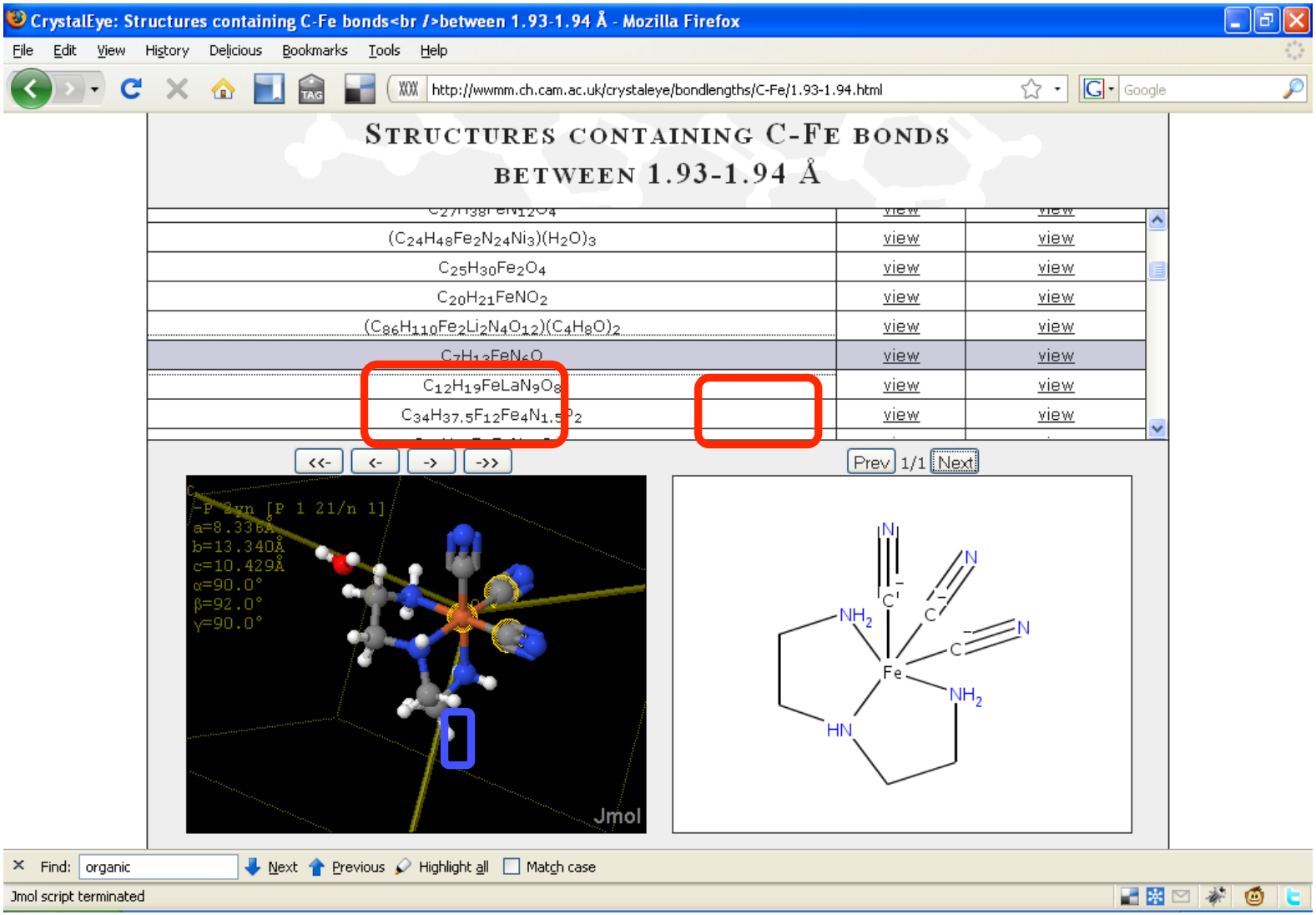
- 

- n

Dissemination

1. Browse
2. Search
3. RSS feeds
4. Harvesting
5. **Example services**

Dissemination 5 – Example Services

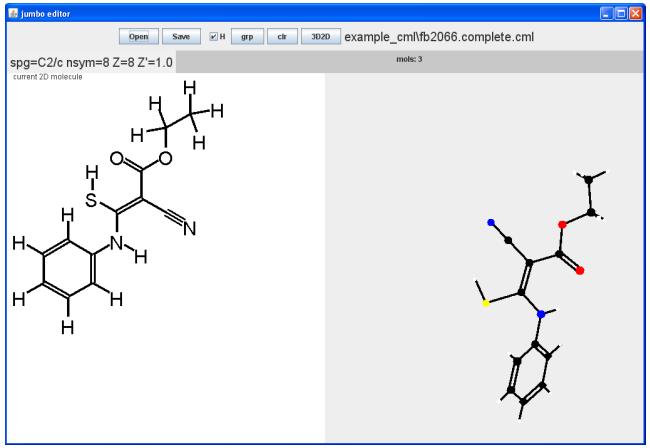


UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

Related Work – C3DE

- 2D Structure/connection table editor – sponsored by IUCr .
- Near-automatic authoring of 2D structures from CIF.



Thanks

- UCC – Peter Murray-Rust, Jim Downing, Andrew Walkingshaw
- Summer students – Mark Holt, Dan Hagon, Lee Harper, David Bebb
- Collaborators – IUCr, University of Southampton
- Sponsors – IUCr, JISC

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics