A Tale of COD

Saulius Gražulis
Kyoto 2008

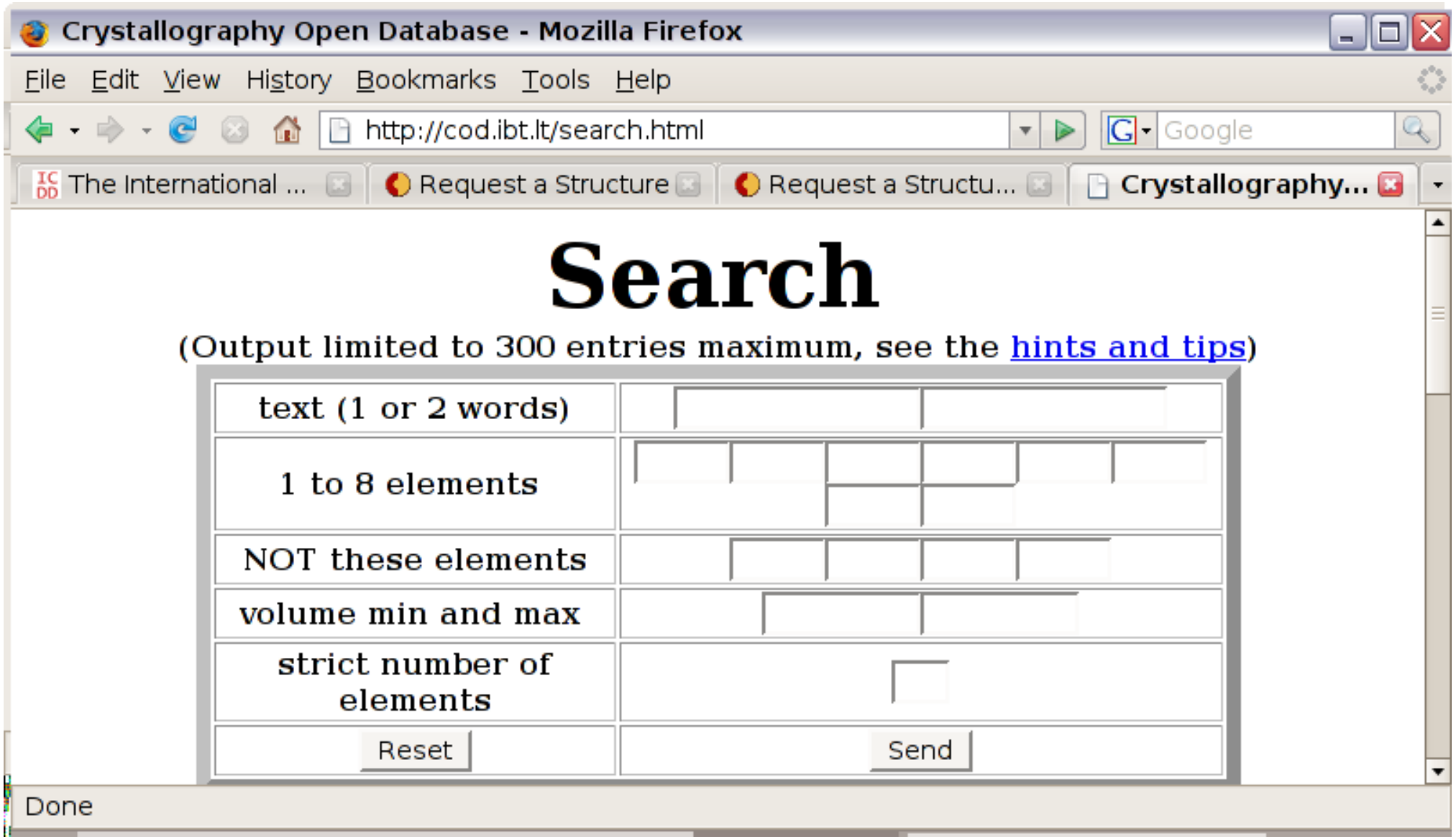
Crystal structure databases

- PDB, NDB
  - open access
  - free to be copied
  - serve as base for numerous derived databases
- CSD, ICSD, ICDD/PDF, CRYSMET
  - proprietary, subscription based
  - copying is not permitted
  - contrast the situation with PDB...

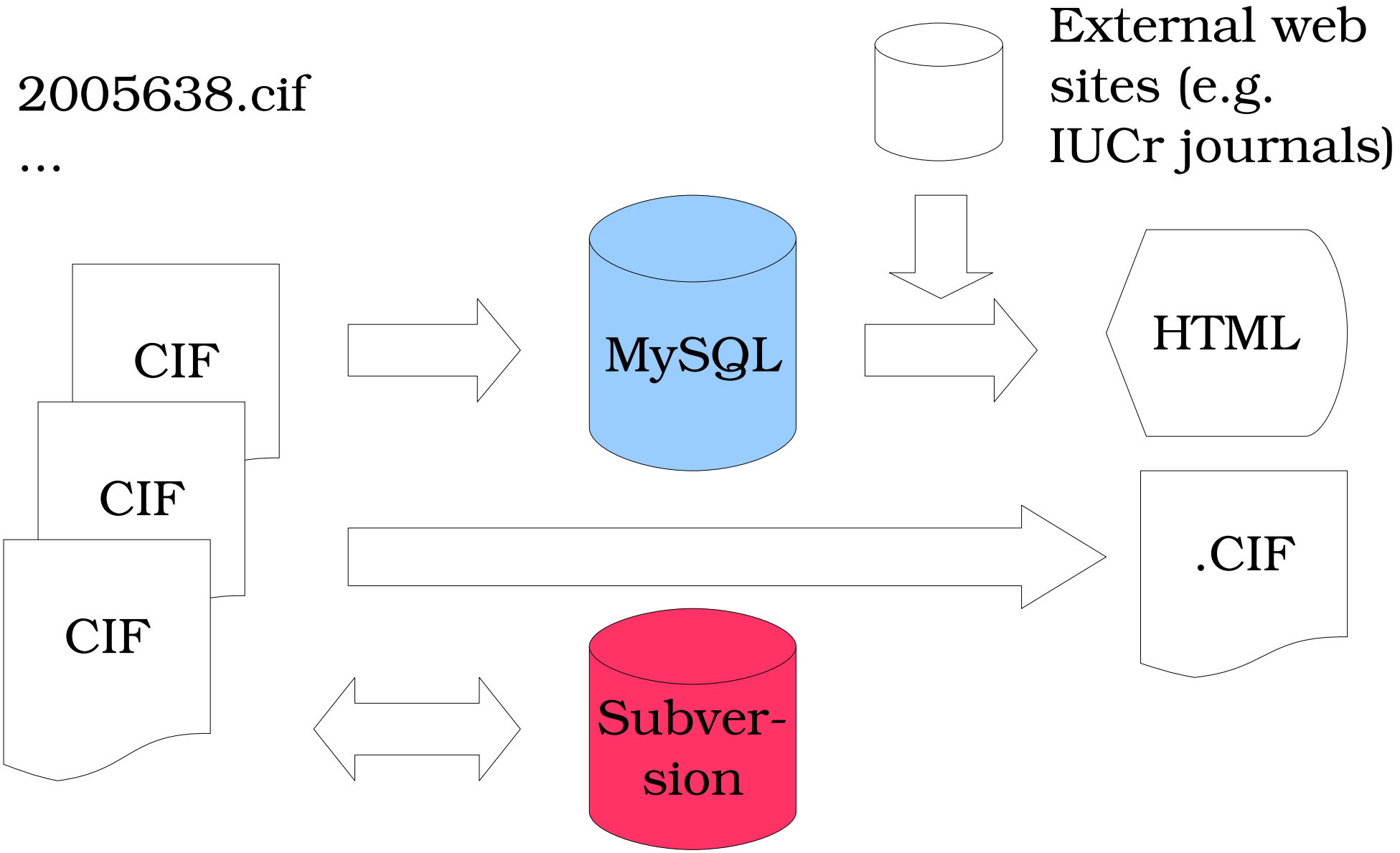
The COD way

- “All data on this site have been placed in the **public domain** by the contributors” (http://www.crystallography.net)
- Updated daily: 71250 entries in the COD
- Collect published structures from peer-reviewed journals and donations by well-established crystallography labs
- Provide high quality data: syntax clean CIF files; increasingly stringent validation tests

COD search interface



COD organisation



COD data sources

- Donations:
  - Mineralogical Society of America
  - Mineralogical association of Canada
  - Laboratoire de Cristallochimie et Physicochimie du Solide
  - Laboratoire de Cristallographie et Sciences des Matériaux CRISMAT
  - Laboratoire des Oxydes et Fluorures, Institue de physique del la Matière Condensée
- Donations by journals:
  - IUCr journals
- Collections by volunteers from the peer-reviewed journal supplementary data.

COD contents

- Structure data in CIFs, 1 structure/CIF
- Crystallographic data & coordinates (of course ;)
- Bibliography, chemical information, back-reference (original file)
- Empty, unrecognised, irrelevant, copyrighted tags are excluded

COD data deposition & quality checks

- Check syntax
- Check semantic consistency
- Check duplicates
- Split structures into separate files
- Add missing information (bibliography, etc.)
- Insert into subversion repository and into the SQL database

Problems with published data

- Impression: ~40% of published CIF files contain syntax errors; ~1% in the IUCr journals in recent years contain semantic problems...
- “These [i.e. syntactic and missing data] problems, which affect about 40% of incoming CIFs” (Frank H. Allen, The Cambridge Structural Database: ..., Acta Cryst. B, 2002, **58**, 380 – 388)

Access to COD data

- COD
  - http://www.crystallography.net
  - http://cod.ibt.lt/
  - svn://cod.ibt.lt/cod
  - rsync://cod.ibt.lt/cif
- PCOD
  - http://www.crystallography.net/pcod
  - svn://cod.ibt.lt/pcod

Community based effort?

- Wiki (Wikipedia) – like review, error correction, annotation, linking with other resources
- Invited editors and reviewers?
- Self-registered editors and reviewers?

COD applications

- Source of ligands for macromolecular crystallographers
- Collecting statistics (representative data subset?)
- Search-match software
- Teaching
- Software testing and validation

COD financing sources

- Volunteers and contributors
- COD Advisory Board
- Lithuanian Research Council
- Government funding?
- Private granting agencies?
- Company donations?

COD prospects

- Docking
- Software testing (after inclusion of Fobs data)
- Crystallographic publication validation and review
- Rational drug design
- QSAR
- Materials research
- Semantic web

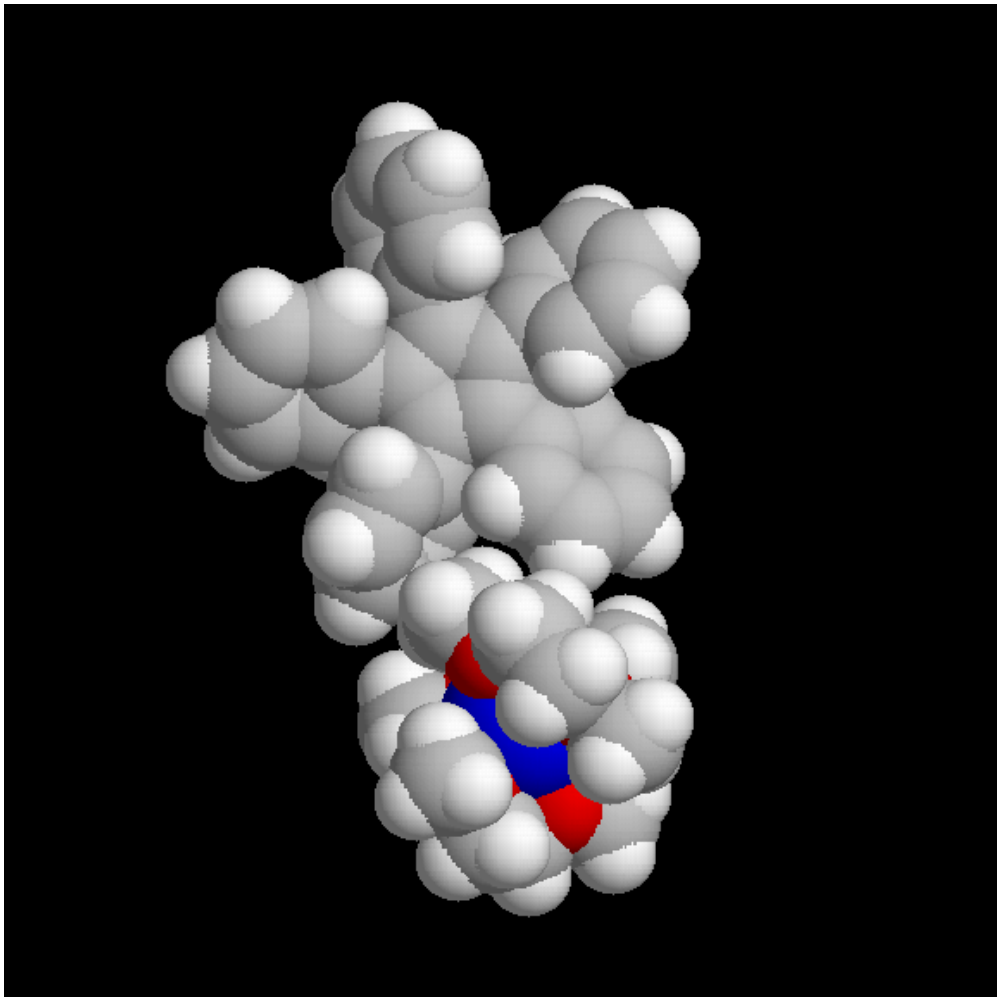
Data to knowledge

- Database is not yet knowledge...
- ... but it is nowadays an important prerequisite!
- Semantic webs?
- Automatic inference?

Acknowledgements

- Volunteers – data collectors
- Advisory board: Daniel Chateigner, Xiaolong Chen, Marco Ciriotti, Robert T. Downs, Saulius Gražulis, Armel Le Bail, Luca Lutterotti, Yoshitaka Matsushita, Peter Moeck, Miguel Quirós Olozábal, Hareesh Rajan, Alexandre F.T. Yokochi
- Special thanks to:
  Elena Manakova, Justas Butkus, Patrick Ducrot
- Lithuanian Science Council Student Research Fellowship Award
- IUCr for permission to automatically download the published CIFs

Questions?



Copyright issues

- Copyright covers works of authorship (novels, verse, sci. papers, computer programs)
- Copyright covers **only** the expression of ideas
- Copyright **does not** cover:
  - Ideas
  - (scientific) facts
  - Simple forms (i.e. ones that do not contain individual's “trace of the hand”)

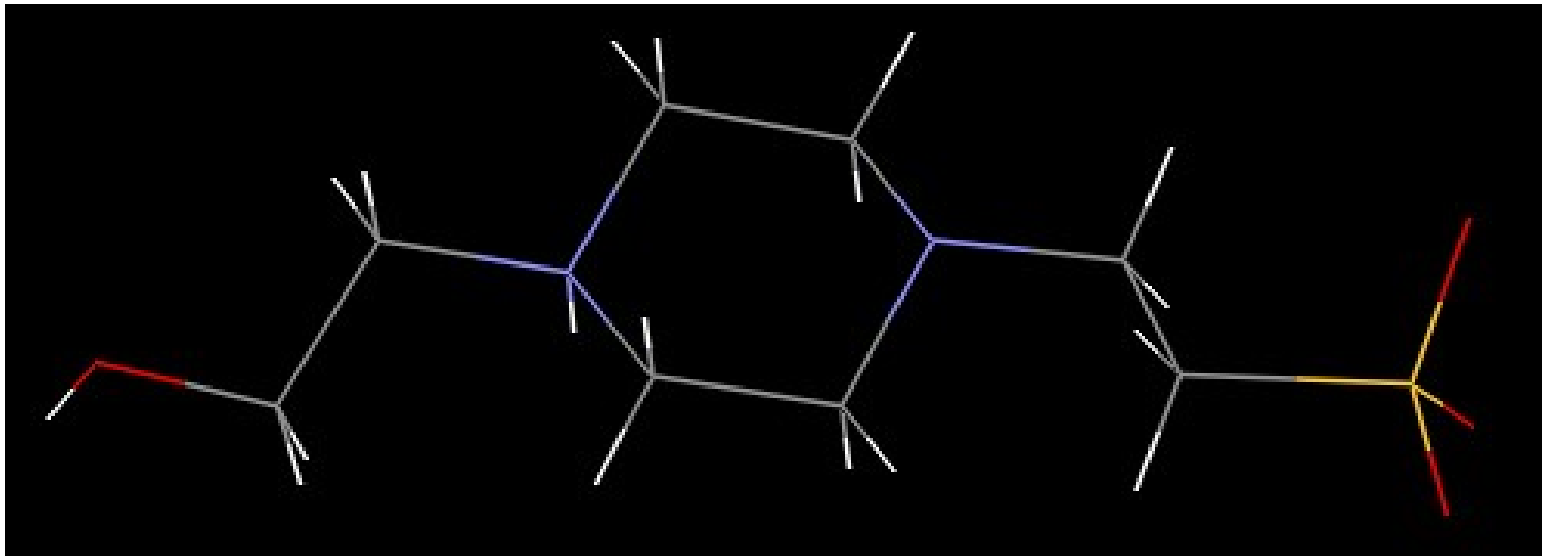
COD copyright policy

- Include data:
  - _atom_site_fract_x 0.333
- Exclude potentially copyrighted text:
  - _publication_text
    ;
    Introduction

    We have solved ...
    ;

The problem?

- Model glycerol/HEPES/MES/Tris into my protein structure:
  - need ideal(ised) coordinates:



Where are my (epps, HEPES) coordinates?

- Sources of coordinates:
  - Quantum mechanical calculations
    - tricky and time consuming
    - need verification
  - Idealised geometry
    - only well-known compounds
    - need verification
  - **X-ray diffraction experiment**
    - precise and accurate
    - time consuming, but >500 000 molecules published
    - **need access to published experimental data**

Crystallographic databases

- Structural information is scattered in several databases:
  - CCDC for organic molecules
  - ICSD for inorganic
  - ICDD/PDF for powder data
- All these databases are proprietary, subscription based “products”
- Contrast the situation with PDB or NDB...

Obtaining data from CCDC?

- "... CCDC provided a web form for data retrieval, which **requires** you to enter brief literature **citation** details and the CCDC **Deposition Number** (CCDCnnnnnn) which should appear in the paper"
  (http://www.ccdc.cam.ac.uk/products/csd/request/)
- Individual CIF data sets are provided ... on the understanding that they are used for bona fide research purposes only. They ... **may not be copied or further disseminated in any form**
  (http://www.ccdc.cam.ac.uk/products/csd/request/request.php4)