Extracting data, information, and knowledge from an ELN

Colin Bird\textsuperscript{1}, Simon Coles\textsuperscript{1}, Aileen Day\textsuperscript{2}, Jeremy Frey\textsuperscript{1}, Richard Whitby\textsuperscript{1}

\textsuperscript{1} Chemistry, University of Southampton
\textsuperscript{2} Royal Society of Chemistry, Cambridge
The Moving Finger writes; and, having writ, Thinks on ...
Inputs ... e.g.

Metadata

ELN

Plans

Push

Pull
Output ➔ Curation of input

• Metadata!
  – Facilitates discovery
  – Avoids any need to know structure/organisation
  – Provenance

Why do we need access?
Dial-a-Molecule

• UK Grand Challenge Network
• Step change: *how can we make molecules in days not years?*
• Essential: exploit the vast body of currently inaccessible chemical data and information held in ELNs

http://www.dial-a-molecule.org/wp/roadmap/

• Example: Reaction pathway selection
Pulling data from ELNs

• Does the ELN hold the information I need?
• If so, what is its provenance?
• In what format(s) is the data available?
• What do I need to know about access control and/or licensing conditions?
• How do I obtain further information?
• ...

Dial-a-Molecule working party

First steps towards semantic descriptions of ELN records – DOI: 10.1186/1758-2946-5-52
Extraction process

Detail

Knowledge layer

Information layer

(Processing) Data layer

Wisdom?

Encounter

Explore

Extract

Time
Three-layer model: perspectives

• Entry (Encounter): Knowledge layer
  – Discover what data and information are available
  – High-level description; access indicators

• Explore: Information layer
  – Nature & context of content (analytical example)
  – Expect to be in ELN (or equivalent)

• Extract: Data (Processing) layer
  – Data (raw data) for reuse or repurposing
  – Might be in a data warehouse
Do the layers overlap more than we might wish?
Working group

Use cases

Schema

Standards

Processes (Protocols)

Use cases:
- Map from range of communities;
- Share data models;
- Test models against use cases;
- Darwinian?

Schema:
- Many options.
- BatchML?
- ExptML?
- Possibly develop container schema.

Archives in format accessible 25 years on

Standards:
- Support and/or develop;
- Once they exist, apply pressure to use them

Processes:
- Make it easy for chemists in the lab to capture metadata; make process easier to do than not to do
Animated text from previous (Working group) slide

**Use cases:**
Map from range of communities; share data models; test models against use cases; Darwinian?

**Schema:**
Archive in format accessible 25 years on

**Schema:**
Archive in format accessible 25 years on

**Processes:**
Make it easy for chemists in the lab to capture metadata; make process easier to do than not to do