The Problem:
Vast quantities of chemical data (e.g. crystal structures, NMR spectra, experimental reports) are generated every day. The majority of this data is never published, and the data that is published is fragmented, trapped in legacy formats and difficult to discover.

The Solution:
Semantically Enriched Linked Open Chemical Data: browsable, searchable, discoverable and interpretable by humans and machines alike, using standardized extensible data formats (Chemical Markup Language) and technologies (HTTP, RDF).

Architecture:
Chem#-ingests data from a variety of sources:
- Atom Feeds
- SWORD
- Web Spiders

Data is semantified, and converted into CML and RDF representations. Normalisation and validation operations may be performed.

Data is stored as linked CML and RDF in a Chem# repository. Data items can be organised into aggregations and collections.

Data is published to humans via webpages, downloads and search forms.

Data is also discoverable and searchable by machines using RDF and SPARQL.

Collections of data can be monitored by subscribing to feeds.

Subsets of data can be explored using visualisation tools such as Pivot.

CLARION

CrystalEye

ACS
Nature
RSC
IUCr

Aggregate reports, such as bond-length distributions and data feeds are generated.

Over 200,000 published crystal structures are searchable and downloadable as semantic chemical data!

The Vision:
Individuals who cannot host their own repository publish data into the cloud

Subject-specific repositories aggregate data from upstream repositories

Chemist’s machine agent subscribes to repositories and automatically notifies chemist about new data and publications relevant to their work!

Other repositories store data collected from crawling publications and legacy data

Publishers host repositories linking semantic chemical data to publications

Scientists publish semantic chemical data into domain-specific institutional repositories

Repositories share data via Atom/RSS feeds

Repositories share data via Atom/RSS feeds