Merging human-readable and computer-readable structure data representations into unified documents

Wolf-D. Ihlenfeldt, Königstein, Germany

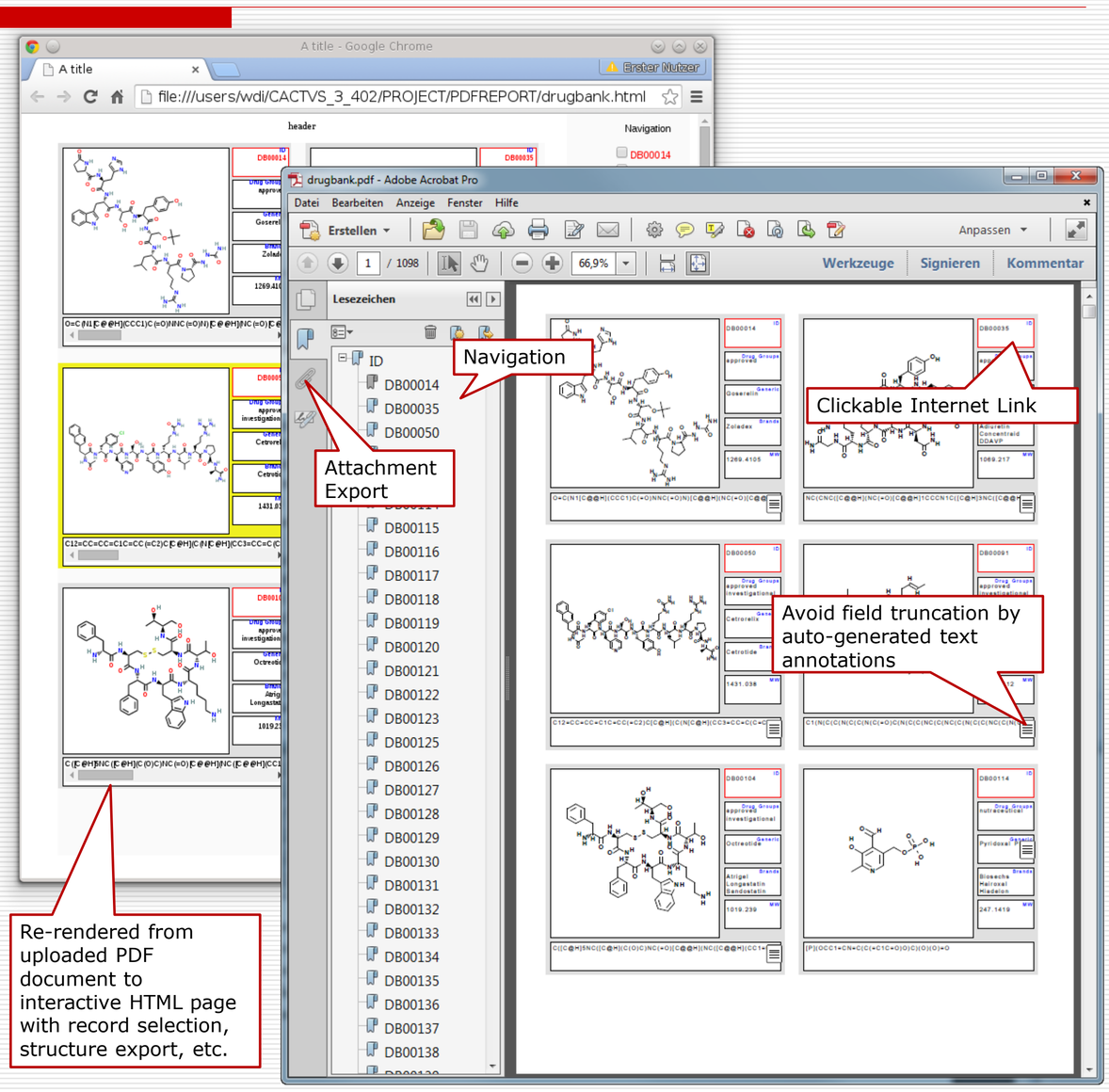
Packaging and Distributing Chemistry Datasets

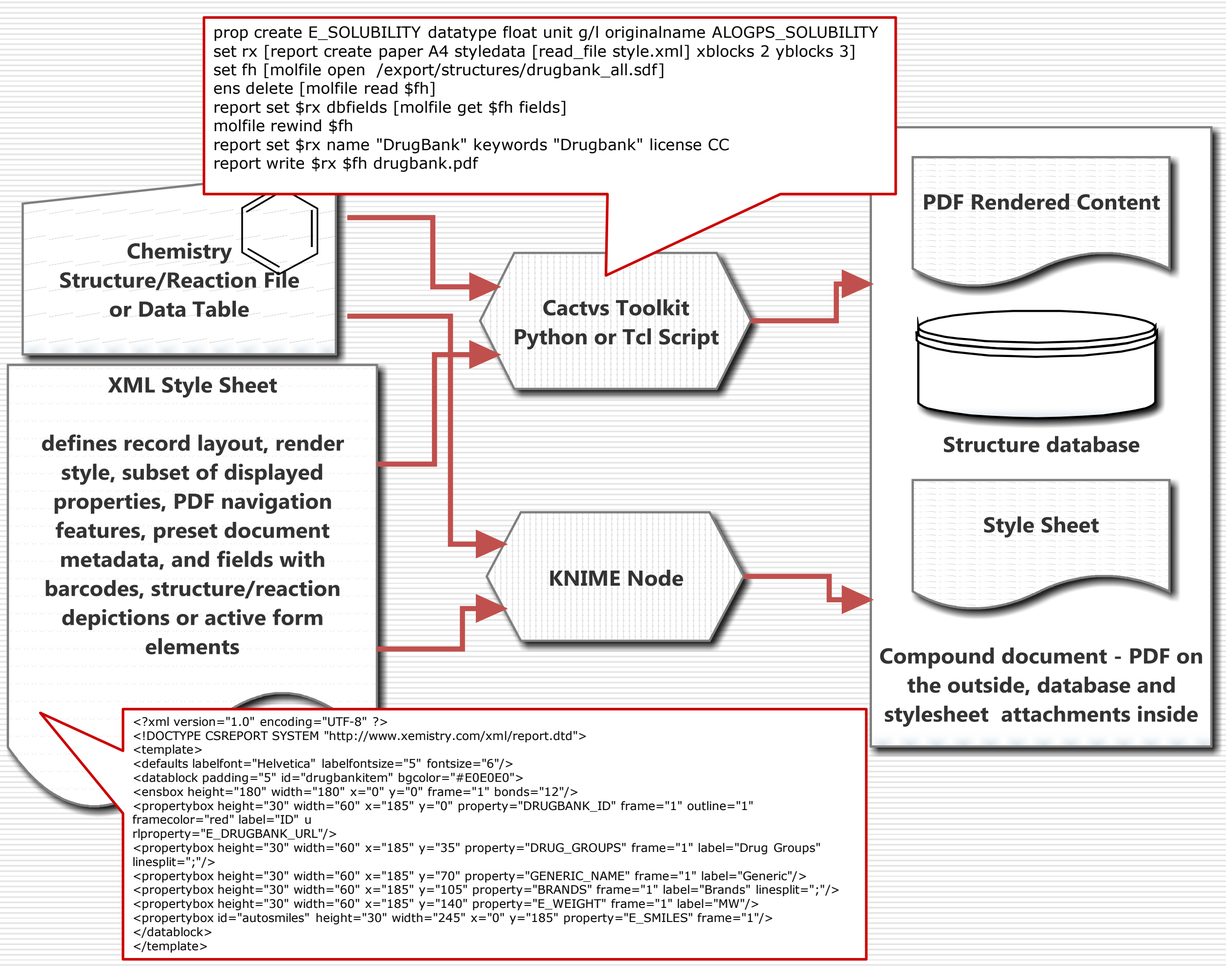
A media gap continues to exist between chemical information designed for human consumption (text and structure/data renderings, nowadays mostly distributed as PDF files) and raw supporting data, which is normally provided separately in formats such as SD files, SMILES, InChI, or numeric tables in CSV or Excel format.

These are separate realms – neither can structures be extracted from a PDF without resorting to unreliable, exotic tools like Chemical OCR, nor can anybody obtain a general overview about the contents of a structure file without expert viewer software usually not present on readers such as a Kindle or iPad which have become the standard means of perusing chemical literature.

We have developed a tool chain to merge the technologies of chemistry-aware SQL relational databases (storing raw data, structures and reactions) and PDF files with visual data representations to address this issue.

The result is a unified medium for chemical data exchange packing the triad of a database, a renderer/layout style sheet, and rendered PDF content into a single Smart PDF file. This file can be viewed (or printed) with any standard PDF viewer. Simultaneously, its data load can be queried, analyzed, exported in original or converted formats or re-rendered with different layouts or information content as a whole or in a user-selected subset by means of a Web portal, user-designed toolkit scripts, or pipeline processors such as KNIME.





Operations on Smart PDF files

• Peruse in any PDF viewer (or on paper), from viewer copy&paste printed data, export embedded attachments to disk, inspect dataset metadata. No plug-ins of any kind required.

• Conveniently shared via email and other Internet channels - PDFs do very well at penetrating overzealous email gatekeepers, much better than ZIP or SDF with or without compression.

• Upload PDF file to public or private Smart PDF processing Web site. These provide functions such as subset selection by data value or structure query, structure export with format conversion, clipboard transfer to ChemDraw/SymyxDraw, or data-preserving re-rendering with a modified style sheet.

• Process the PDF file with the Cactvs Cheminformatics Toolkit. Use custom scripts to read, write, filter and modify the PDF rendering or its data payload. Anything which can be done with an SDF or data table file can be done with the PDF file as only input object, by means of transparently accessing its embedded database, without danger of getting out of sync between raw data and readable report.

• Even as plain PDF or on paper, these are not dead documents. Interactive elements can still be present – for example embedded barcode or QR codes of properties such as URLs, PDF form elements for expert surveys, or Smartpen patterns (under development).

FAQ

▪ Is this a proprietary PDF variant?

No. Everything in these PDFs is fully standards-compliant- We are just exploring some of the less travelled corners of the specification. A Smart PDF can be opened with any PDF reader, and the database or style sheet exported from any viewer which knows about in-file attachments (most do, and Acrobat Reader certainly does).

▪ But the embedded database is something proprietary?

No. It is a standard portable SQLite database file. These are directly usable on 32 and 64 bit platforms independent of byte order. It can be processed with the standard SQLite tools and libraries – structure queries obviously excluded.

▪ So how do I run structure queries on it without buying into your complete software stack?

We offer a chemistry cartridge for SQLite. It can be dynamically loaded into any SQLite-enabled application. It adds the usual structure query functionality (full-structure, substructure, superstructure, similarity. etc.), as well as property computation, format conversion etc. as additional functions for use in standard SQL statements.

Isn’t it slow and bloated?

No. This is actually a very feasible technology for datasets up to the 100K structures / 5M data points range.

Test case: DrugBank (www.drugbank.ca) full dataset, 6500 hydrogen-stripped structures, 25 SD data fields.

Size of original SDF: 22 MB
Size of Smart PDF: 27 MB - graphical structure and data renderings, PDF navigation, smart data annotations, Web links, the embedded database with all structures, the complete SD field set and structure query accelerator columns, plus the embedded style sheet file all included. Admittedly, a compressed SDF would be smaller.

Acrobat reader open time, standard PC: <1s
PDF and database creation from raw SDF: 50s
Re-rendering with different style sheet and data selection: 10s
Extracting structure database file from PDF: <1s
Database queries (full-structure, substructure, similarity, etc.): <1s
Full read-back of all structures and data from PDF into Cactvs toolkit: 2s
Substructure-based selection and export of 100 structure subset as SDF: 2s