

# Oral Roberts University

---

From the Selected Works of Andrew Lang

---

July, 2009

## Beautifying Data in the Real World

Andrew Lang

Jean-Claude Bradley, *Drexel University*

Rajarshi Guha, *National Institutes of Health*

Pierre Lindenbaum

Cameron Neylon, *University of Southampton*, et al.



This work is licensed under a [Creative Commons CC\\_BY-SA International License](https://creativecommons.org/licenses/by-sa/4.0/).



Available at: <https://works.bepress.com/andrew-sid-lang/>

## CHAPTER SIXTEEN

# Beautifying Data in the Real World

*Jean-Claude Bradley, Rajarshi Guha, Andrew Lang, Pierre Lindenbaum, Cameron Neylon, Antony Williams, and Egon Willighagen*

### The Problem with Real Data

**THERE ARE AT LEAST TWO PROBLEMS WITH COLLECTING “BEAUTIFUL DATA” IN THE REAL WORLD AND** presenting it to the interested public. The first is that the universe is inherently noisy. In most cases collecting the same piece of data twice will not give the same answer. This is because the collection process can never be made completely error-free. Fluctuations of temperature, pressure, humidity, power sources, water or reagent quality, precision of weighing, or human error will all conspire to obscure the “correct” answer. The art in experimental measurement lies in designing the data collection process so as to minimize the degree to which random variation and operator error confuse the results. In the best cases this involves a careful process of refining the design of the experiment, monitoring size and source of errors. In the worst case it leads to people repeating experiments until they get the answer they are expecting.

The traditional experimental approach to dealing with the uncertainty created by errors is to repeat the experiment and subject the results to statistical analysis. Examples of repetition can be found in most issues of most scientific journals by looking for a figure panel that contains the text “typical results are shown.” “Typical results” is generally taken to mean “the best data set we obtained.” Detailed statistical analysis, although in principle a more rigorous approach, can also be controversial and misleading. Arguments often rage

in the comments pages of medical journals over the appropriate approaches to take to remove confounding correlations from the analysis. The links between skepticism about “typical” results and arguments over statistical approaches is a lack of access to the raw data. If the underlying data were available, then people could simply do the analysis and check it themselves. This would likely not reduce the number of arguments, but would at least mean they were better informed.

The second part of the problem is that, until recently, space limitations in print journals have limited the amount of data that can be presented, making it difficult or impossible to present the whole body of data and analysis that supports the argument of the paper. However, in a world where publishing has moved online, this is no longer a viable excuse. It is possible to present the entire data set on which an argument is based, at least in research where data volumes are in the kilobyte to gigabyte scale. There is therefore a strong argument for presenting the whole of the data. This, however, raises the problem of how to present data that may be inconsistent, that may include mistakes, but nonetheless presents the whole picture of how a conclusion was reached. In short, the question is how to show the beauty that lies under the surface of the data in a clear way, while at the same time not avoiding or hiding the blemishes that may lie on the surface.

We believe that the key to successfully reconciling these apparently conflicting needs is transparency. Providing the raw data in as comprehensive a fashion as possible and a full description of all the processing and filtering that has taken place means that any user can dig down to the level of detail he or she requires. The raw data will often be difficult or impossible to present in a form that is naturally machine-readable and processable, so the filtering and refinement process also involves making choices about categorization and simplification to provide clear and clean datafiles that can be repurposed. Here we describe the approach we have taken in “beautifying” a set of crowdsourced data by filtering and representing the data in an open form that allows anyone to use it for his own purposes. We show the way this has enabled multiple researchers to prepare a variety of tools for visualization and analysis, creating a collaborative network that has been effective in analyzing the results, suggesting further experiments, and presenting the results to a wider audience in a way that traditional research communication does not allow.

## Providing the Raw Data Back to the Notebook

As part of a wider program of drug discovery research (Bradley 2007) led by Professor Jean-Claude Bradley, we wished to predict the solubility of a wide range of chemicals in nonaqueous solvents such as ethanol, methanol, etc. Of greatest interest was the solubility of aldehydes, carboxylic acids, isonitriles, and primary amines—components required for the Ugi reaction that the Bradley group use to synthesize potential antimalarial targets (Bradley et al. 2008). The solubility of a specific compound is the quantity of that compound that can be dissolved in a specific solvent. Building and validating a model that could predict solubility would require a large data set of such solubility values. Surprisingly, there was no readily available database of nonaqueous solubilities. We therefore elected to crowdsource the data, opening up the measurements to anyone who wanted to

be involved (<http://onschallenge.wikispaces.com/>). However, this poses a series of problems. As anyone can contribute measurements, we have no upfront way of checking the quality of those measurements.

The first stage in creating our data set therefore required the creation of a detailed record of how each and every measurement was made. The measurement techniques, precision, and accuracy of different contributions all vary, but all the background information is provided in human-readable form. This “radical sharing” approach of making the complete research record available as soon as the experiments are done, called Open Notebook Science ([http://en.wikipedia.org/wiki/Open\\_Notebook\\_Science](http://en.wikipedia.org/wiki/Open_Notebook_Science)), is not common amongst professional researchers, but it is a good fit with our desire to make a complete and transparent data set available. We utilize a Wiki, hosted on Wikispaces (<http://onschallenge.wikispaces.com>) to hold these experimental records, and other services such as GoogleDocs (<http://docs.google.com>) and Flickr (<http://flickr.com>) to hold data (Figure 16-1).

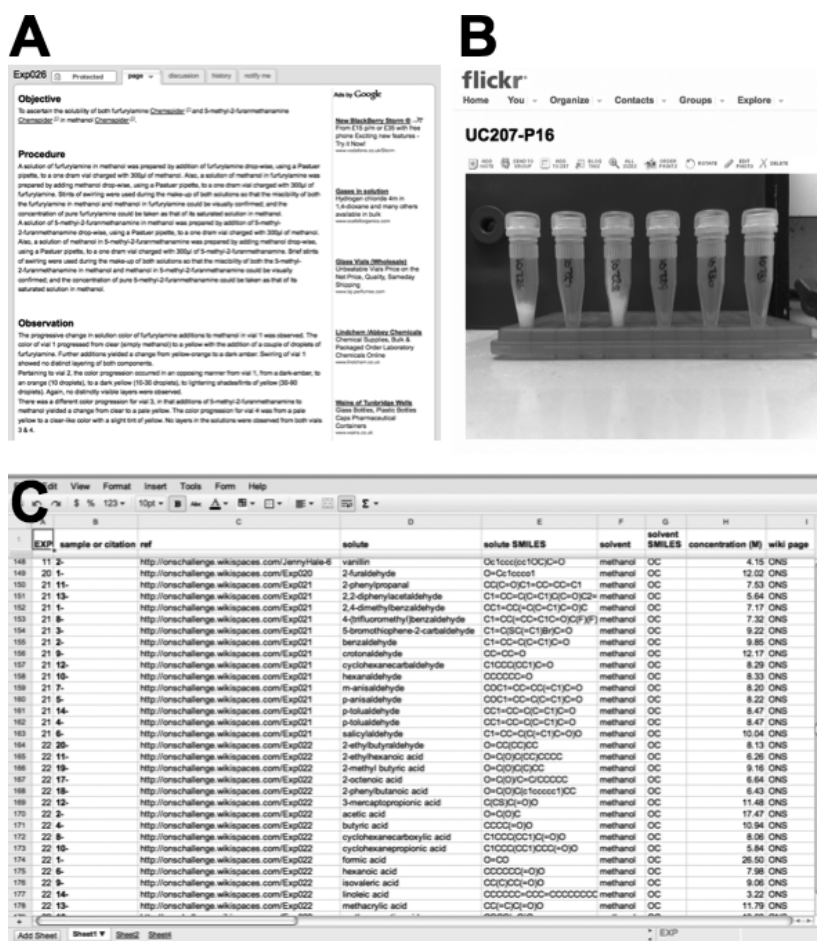


FIGURE 16-1. Using free generic services to host the record of experimental work and processed data. (A) Part of the page of a single experimental measurement. (B) Images taken of the experiment hosted on Flickr. (C) A portion of the primary data store on a GoogleDocs spreadsheet. (See Color Plate 53.)

The actual database of values extracted from the experimental descriptions is stored in a GoogleDocs spreadsheet to generate the primary aggregated record of the project. Each and every measurement is presented, along with a link back to the primary record. This link is crucial whether the data is being read by a machine or by a human, as it provides both the provenance of the measurement (i.e., who is making the assertion) and the record of evidence for that assertion. A human can click through to check how the measurement was made, and a machine reader can download or scrape the record if desired.

In a sense the spreadsheet is the first point at which the record from the lab notebooks is converted into data that can then be criticized and filtered. The choice of a GoogleDoc may appear an idiosyncratic choice from a technical perspective, but it is based on a number of requirements: functionality that enables us to present and share the data in its primary form; an interface that is familiar to experimental scientists and requires a minimum of additional work on their part; and free and hosted services, maintained by a large stable company, enabling anyone in the world to replicate this information-processing model with minimal effort. Finally, there is the ability to access the data via a powerful and flexible API. Very few other approaches both enable the average scientist to work with, add to, and download the primary data in a form that is familiar, and also provide powerful programmatic access to the underlying data.

## Validating Crowdsourced Data

As data is collected by different researchers using different methods, incompatible values are likely to arise. These may appear as outliers or simply as a wide spread of results. Traditionally, with no additional information, researchers had little choice but to give equal weight to each measurement or apply statistical methods to exclude outliers. But, as we have adopted an Open Notebook approach requiring the full record of how each measurement was carried out, each measurement can be evaluated in the context of the information recorded. In several cases this allows a scientist familiar with the methods reported to exclude questionable data points on the basis of inappropriate conditions or a failure to report an important parameter.

In the case of solubility, mixing time and evaporation conditions proved to be important factors. A good example of this was the determination of the solubility of 4-nitrobenzaldehyde in methanol. Of the five measurements taken, three are significantly lower than the other two (Figure 16-2, shown later in this chapter; <http://oru.edu/ccda/sl/solubility/ugidata.php?solute=4-nitrobenzaldehyde&solvent=methanol>). This method was based on preparing a saturated solution of 4-nitrobenzaldehyde in methanol, evaporating the methanol, and then weighing the residue left behind. It is crucial that a fully saturated solution is prepared, and this was generally done by adding solute with mixing until visible solid remained in the tube. By examining the detailed record of the experiments, it is clear that the three lower values are from experiments where the solutions were mixed only briefly. The two higher measurements are from experiments where the mixing was carried out over several hours, showing that extended mixing is required (<http://usefulchem.blogspot.com/2008/12/mechanical-turk-does-solubility-on.html>).

The availability of the raw experimental record enables all researchers both to identify those measurements that are doubtful and to benefit from the experience of these “failed” experiments. This is the nature of scientific research. A balance has to be struck between recording the details of an experiment and being efficient. Often the purpose of initial experiments is for the researchers to identify what factors are important to pay attention to. Unfortunately, this information is generally not shared with the research community.

Rather than exclude results that are doubtful, we have opted to tag the measurements (with “DONOTUSE”) and provide a reason for exclusion. This allows other researchers to click through to the original lab notebook pages and evaluate the raw data for themselves. Errors can occur everywhere—including during the validation process—and “incorrect” values may be useful for some purposes. Full transparency makes it possible for each user to decide what values she wishes to include in her analysis. It also reduces (but does not eliminate) the risk that errors remain hidden. An example of this type of markup is demonstrated in the reporting of the measurements for the solubility of vanillin in methanol (<http://usefulchem.blogspot.com/2008/11/what-is-solubility-of-vanillin-in.html>).

There is clearly a gray area here between those values we have marked as doubtful or untrustworthy and those that we have left unmarked despite potential issues or areas of disagreement. At the end of the day these are matters of scientific judgment, and there is much opportunity for disagreement. The primary record of the experiments remains available in every case and can be examined. In addition, the history of the spreadsheet is available and can also be examined. A balance needs to be struck between providing a useful data set and the degree to which every decision and mistake can be presented. This is a challenging balance to get right.

## Representing the Data Online

Our aim is to make all of the experimental record and processed data available online. This raises a number of issues for how to represent the data in a useful form on the Web, including the choice of standardized identifiers, visualization tools, and approaches to data integration.

### Unique Identifiers for Chemical Entities

To make our data useful, it is important that the chemical entities be described using a recognized standard. Without this, integration with other data sets will be difficult or impossible. In chemistry, some would argue that CAS Registry Numbers ([http://en.wikipedia.org/wiki/CAS\\_registry\\_number](http://en.wikipedia.org/wiki/CAS_registry_number)) would be ideal for identifying chemical entities. However, CAS numbers are proprietary in nature, cannot be converted to the chemical structure, are a lookup only, and are dependent on an external organization to issue. We would prefer identifiers that are open in nature, freely available for exchange, and can be converted to and from a chemical connection table.

The IUPAC International Chemical Identifier (InChI, pronounced “INchee”) provides a non-proprietary standard and algorithms along with supporting open source software (<http://en.wikipedia.org/wiki/Inchi>) that enable the generation of identity strings that can be converted

back to structures (see <http://www.qsarworld.com/INCHI1.php> for a recent review). InChI is gaining significant support as a standard across software vendors, publishers, and developers. The problems with the algorithm—which mean it is possible to generate multiple InChIs for a single structure—are being addressed by the development of the Standard InChI. For some purposes the InChIKey, a hash of the InChIString, is useful, but this cannot be converted to a structure and must be used via a lookup table to *resolve* the chemical structure.

SMILES (<http://en.wikipedia.org/wiki/SMILES>) is a common format for representing chemical compounds, providing strings that are quite compact and can be converted to and from chemical structures. However, there are multiple forms and implementations of the SMILES algorithm, leading to multiple SMILES for the same entity. We are currently using SMILES in this work due to their simplicity and ease of searching. As noted in “Enabling Data Integration via Unique Identifiers and Self-Describing Data Formats” on page 269, it is possible to convert our SMILES to InChIs, automatically enabling us to integrate our data into the growing web of data represented by this preferred identifier.

### Open Data and Accessible Services Enable a Wide Range of Visualization and Analysis Options

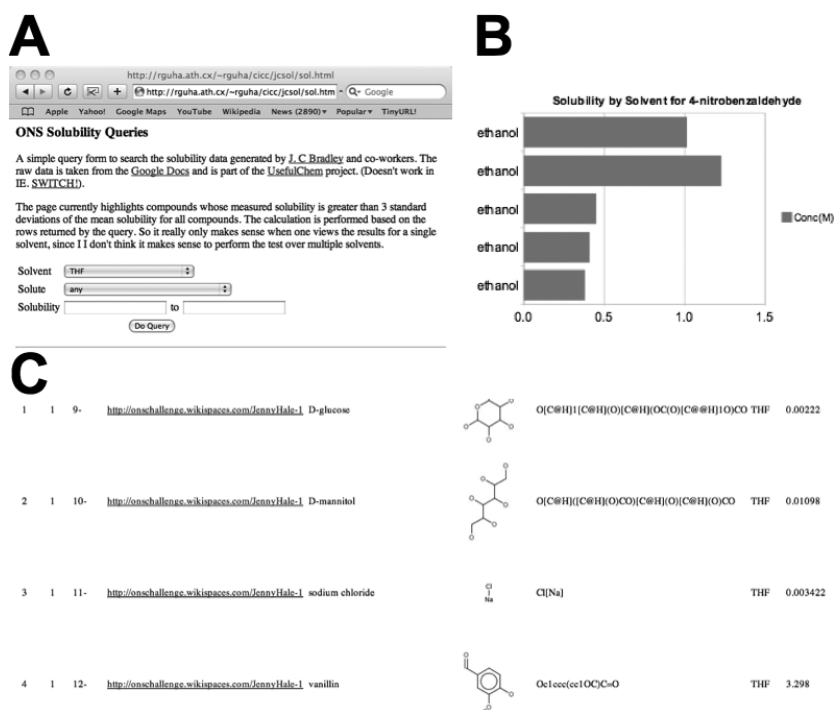
Given a standardized, free, and accessible storage infrastructure for the primary solubility data, the next step is to analyze the data. Analysis could range from simply generating a summary of which measurements have been carried out to complex statistical representations of models derived from the data. In either case, it is necessary to access the data for processing by automated routines. It is also desirable to be able to use other sources of information to enrich the data. This in turn requires a recognized standard to be used in the primary representation of the data.

The spreadsheet contains a number of columns, with each row being the record of a single measurement. Both solvent and solute are represented in two different forms: a human-readable common name and a SMILES code. As with the choice of GoogleDocs as the primary representation of the data, the use of both human-readable and machine-readable representations is crucial to gaining the most benefit from the data set. The only piece of information that does not require two representations is the numerical representation of the solubility itself.

As has been noted earlier, we made the decision to not remove the most questionable values from the primary data record. This poses a problem for machine readability, as there is no accepted standard approach for saying “this number is a bit dodgy.” For this work we have elected to mark records that are believed to be inaccurate after human curation and to give a reason for the marking. This enables any user to make choices about which records he wishes to include in any analysis, either manually or automatically. For any further analysis, the data, as represented in the spreadsheet, must be accessed. The simplest approach would be to simply export the data, for example, into a comma-separated text file, and analyze it using some external software. However, this would lose the immediate

link to the most up-to-date data. The GoogleDocs API makes it straightforward to generate web-accessible applications while maintaining the link to the “live” data.

As an example of such an application, we have created a web service that allows users to query the data stored in the spreadsheet and obtain a tabular summary (Figure 16-2). The form-based query interface provides a rapid and intuitive mode of access to the relevant data. The page is simple HTML along with JavaScript and does not require any software on the client side. The data is accessed asynchronously from the Google spreadsheet using the API provided by Google. Once the data is retrieved, one can perform a variety of calculations. In this case, we determine the mean and standard deviations of the solubility, which can then be used to dynamically highlight entries that appear to be anomalous.



**FIGURE 16-2.** Visualization tools for examining the solubility data. (A) A simple form-based input uses JavaScript and the GoogleDocs API to generate (B) a graphical representation of the solubility values selected and (C) a tabular output of the data with rendered 2-D chemical structures. The service is available at <http://rguha.ath.cx/~rguha/cicc/jcsol/sol.html>. Note that these and other services described are dynamic and may not give the same results as those shown here for the same query. (See Color Plate 54.)

While the tabular display of subsets of the raw data is very useful, visualizations can be used to effectively summarize the results of queries. The query application employs the Google Visualization API to generate bar charts based on the data extracted from the spreadsheet. Given the ease with which the data can be extracted from the spreadsheet, one can easily generate a variety of visualizations. In our case, a simple bar chart displaying the solubility of a compound in various solvents provides a rapid summary of the results.



Another aspect of the application is that the table contains 2-D depictions of the chemical structures provided via a REST-based service at Indiana University. A SMILES code is appended to the service URL to insert the 2-D image of the structure into an arbitrary web page. Once again, these features require no special software on the client side. This makes distribution of this specific application extremely simple; one simply has to copy the HTML page to another web server.

Although this is a fairly simple application, it highlights the distributed nature of the solution, combining open data with free visualization methods from multiple sources. More importantly, the distributed nature of the system and free accessibility of the data allow experts in different domains—experimentalists generating data, software developers creating interfaces, and computational modelers creating statistical models—to easily couple their expertise. The true promise of open data, open services, and the ecosystem that supports them is that this coupling can occur without requiring any formal collaboration. Researchers will find and use the data in ways that the generators of that data never considered. By doing this they add value to the original data set and strengthen the ecosystem around it, whether they are performing complementary experiments, doing new analyses, or providing new services that process the data. And, all the time, the link back to the original record can be maintained.

### Integrating Data with a Central Aggregation Service

A valid criticism of our approach is that if it is widely taken up, it will lead to the presence of many disparate and disconnected data resources. Although it is technically feasible to aggregate such resources together using search tools, it remains the case that the researchers usually use a small set of preferred services as their route into the data they are interested in. The gold standard of curated data sets of chemical information is the Chemical Abstracts Service (CAS) maintained by the American Chemical Society. The CAS Registry contains over 40 million substances (<http://www.cas.org/newsevents/connections/derivative.html>) comprised of data extracted from publications, patents, chemical catalogs, and, increasingly, online data sources such as ChemSpider (<http://www.chemspider.com/blog/cas-chemspider-connectivities-and-unintended-collaboration.html>).

ChemSpider is a web-based resource for chemists developed with the intention of “Building a Structure Centric Community for Chemists.” Containing well over 20 million unique chemical entities, sourced from over 150 data sources, ChemSpider has become one of the primary Internet resources for chemists seeking information about chemical entities. For each individual chemical compound, various types of information are associated. This includes different types of identifiers (systematic names, trade names, registry numbers, multilingual names), predicted physicochemical properties, and links to a wide variety of experimental physical, chemical, and spectral data from a wide range of sources (<http://www.chemspider.com/DataSources.aspx>). ChemSpider can therefore be considered as a structure-based link farm to other resources. ChemSpider also provides an environment that allows users to both cleanse and expand the data online. Users can annotate and curate the data,

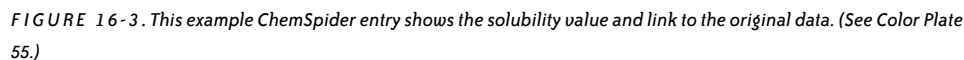
thereby removing erroneous associations with the chemical entities and adding their own information, including links to external resources and other annotations of the data. Users can also deposit new chemical structures to the database and associate spectral data, images, and even video files. A myriad of search capabilities exist, including searching predicted data and structure or substructure-based searches.

ChemSpider therefore provides an ideal environment for connecting to other sources of chemical information where the primary key to the data is the formal identity of the molecule. The combination of providing a central resource for searching for chemical data as well as the deposition of user data makes ChemSpider the logical place through which other researchers would find our data. Although some of the data associated with ChemSpider has been gathered by scraping data from online resources, great care must be taken with such approaches (<http://www.chemspider.com/blog/care-in-nomenclature-handling-and-why-visual-inspection-will-remain.html>), and increasingly data is added only after some form of curation. At present the decision has been made to add specific measurements to ChemSpider on a case-by-case basis to ensure a further human curation step.

The nonaqueous solubility data measured as a part of the ONS-Solubility project is being added to ChemSpider. Currently a few values are available online as supplementary information, presented along with all other physicochemical data that might be available for a specific chemical compound. The solubility data is presented along with a link back to the experimental page, which is consistent both with the ChemSpider approach of acting as a link to the primary data source and our approach of providing a path to the data and then on to the original record. An example ChemSpider record is shown in Figure 16-3. As more nonaqueous solubility values are inserted into the record, either manually or robotically, this data will be exposed and its associated originating source information will be just one click away.

As more data, albeit with differing levels of quality and curation, is made available in the future, it can be expected that this data ingest process will be automated. Efforts are already underway to facilitate this process for the current project, but two major issues will need to be overcome before this process can become widespread. The primary issue is one of trust: which sources of data can be trusted sufficiently to be automatically ingested, and what level of curation should those data sources be expected to have? Is the current level of curation on the primary GoogleDocs spreadsheet adequate, or would a further level of filtering be required? In the current case it seems clear that our approach is not adequate for a service that aims to provide data that chemists can trust without requiring further investigation. As researchers expose more primary data and the interest in automatically ingesting it grows, there will need to be a detailed discussion about when and how data is presented, and what markup is required prior to being placed on a white list. Certification processes can be expected to grow up around the exposure of data, providing a mark of both quality and functionality for exposed data sets.

The second major issue is that of functionality. The number of exposed data sets that are relevant to ChemSpider is currently small, and in most cases reasonably stable, and so



268 CHAPTER SIXTEEN

## Enabling Data Integration via Unique Identifiers and Self-Describing Data Formats

Although the GoogleDoc API provides an easy route toward developing analysis tools and visualization methods that are designed specifically for this data set, it remains the case, even with a unique identifier, that the presentation of the data is not in a standard format. These tools are written against the data set as it is presented in the GoogleDoc spreadsheet. In short, they require a human to understand what descriptors and values are in which column of the spreadsheet. Although it is technically feasible to recognize that a given column contains a SMILES code, it will not be clear to a machine whether this is a solvent or solute, or indeed that the data is about solubility at all. To realize the full promise of connected data (e.g., by supporting automated ingest into ChemSpider and other services), and to provide the data in the most general possible way to other researchers, it is necessary to provide a representation that adheres to a recognized standard in syntax as well as in descriptors.

The Resource Description Framework, or RDF, provides a route toward exposing the data set in a recognized, machine-readable format. With this format, any information is transformed into statements made up of a “subject,” a “predicate,” and a “value.” For example, the fragment shown in the following code states that the object found in the spreadsheet called solute#59 is defined as the resource at the given URL. RDF uses “namespaces,” or sets of recognized concepts, to define relationships between “resources,” where a resource is any object that can be pointed at by a unique identifier. There are four main namespaces used here. The first is the RDF namespace itself, which defines that the file is in RDF and provides other top-level concepts such as “is defined by” or “is a resource.” The second namespace is the spreadsheet containing the data, which is a resource, defined here by the namespace `ons`, which contains specific resources within it, one in each cell of the spreadsheet. The third namespace is Dublin Core (`dc`), which deals with concepts such as name, author, and version. A fourth namespace (`chem`), hosted at <http://rdf.openmolecules.net> (RON), is used to specify that the molecules identified in a specific cell are defined by a specific resource.

As noted earlier, the spreadsheet has its own data *schema*, essentially relying on the fact that each row refers to a single measurement. The GoogleDocs API makes it straightforward to reference a cell using a simple URL. To make this into a declaration in RDF, we need to describe a relationship between the contents of that cell, e.g., “2-octenoic acid” and some other resource. One simple relationship is to identify “2-octenoic acid” as being defined by a specific resource at RON, which again is referenced by a simple HTTP URL. As the contents of the cell are now defined, it is possible to use the external resource to find more information related to that molecule. Resolving the URL will lead to more RDF statements about the same molecule, defined by the service at RON. Similarly, it gives the SMILES and a title to the molecule, all derived from the spreadsheet. For each and every entry representing a molecule in our data set, it is possible to define a standard description and to connect that with other standard definitions, including systematic name, InChI, and SMILES.

```

<ons:Solute rdf:about="http://spreadsheet.google.com/.../onto#solute59">
  <rdfs:isDefinedBy rdf:resource="http://rdf.openmolecules.net/?InChI=
1/C8H14O2/c1-2-3-4-5-6-7-8(9)10/h6-7H,2-5H2,1H3,(H,9,10)"/>
  <chem:inchi>InChI=1/C8H14O2/c1-2-3-4-5-6-7-8(9)10/h6-7H,2-5H2,1H3,
(H,9,10)</chem:inchi>
  <chem:smiles>CCCCC=CC(=O)O</chem:smiles>
  <dc:title>2-octenoic acid</dc:title>
</ons:Solute>

```

Having defined each of the chemical entities found in the spreadsheet, we can now represent each measurement using a piece of RDF similar to that just shown. The RDF defines a new measurement, and gives the solvent, solute, solubility, and the experiment to which this measurement belongs. Again, because we have already defined the identity of each solvent and solute in chemical terms, this measurement information can be linked in and used with any other RDF file that describes data about the same molecule. The fragment shown next uses the XML entity `ons` with the value `http://spreadsheet.google.com/plwwufp30hfq0udnEmRD1aQ/onto#` essentially as an alias to make the XML more readable (&ons;measurement179 is expanded to the full URL with “measurement179” appended).

```

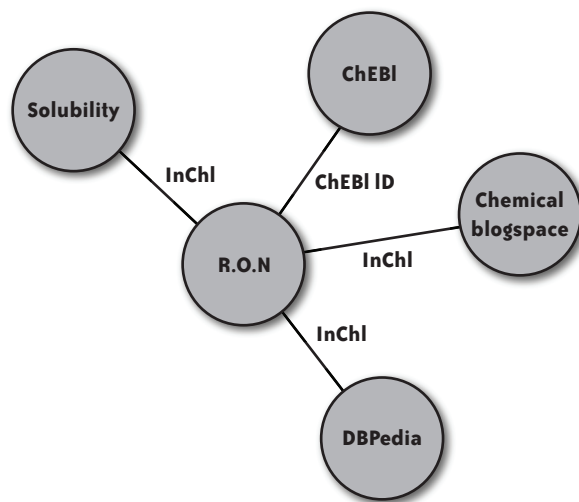
<ons:Measurement RDF:about="&ons;measurement179">
  <ons:solubility>0.44244235315106</ons:solubility>
  <ons:solvent RDF:resource="&ons;solvent8"/>
  <ons:solute RDF:resource="&ons;solute26"/>
  <ons:experiment RDF:resource="&ons;experiment2"/>
</ons:Measurement>

```

These statements, or triples, can then be read or analyzed by any RDF engine and query systems such as SPARQL. By using appropriate namespaces, especially where they are agreed and shared, it is possible to generate datafiles that are essentially self-describing. A parser has been developed (<http://github.com/egonw/onssolubility/tree/>) to generate the full RDF document, available at <http://github.com/egonw/onssolubility/tree/master/ons.solubility.RDF/ons.RDF>. The Chemistry Development Kit (CDK; see <http://cdk.sourceforge.net/>) is used to derive molecular properties from the SMILES, including the InChI. This is a key step: the conversion of experiment-specific information into a datafile that can be read by any system or service that understands RDF. Such services may not necessarily know what to do with specific concepts from new namespaces but will understand how to deal with the categories these concepts fall into, and will be able to parse the data against other resources that use the same namespace.

The real power of RDF arises when multiple resources are linked together via links (see Figure 16-4). It is possible, for instance, to link our experimental data with information in DBPedia, an online resource of information represented in RDF (<http://dbpedia.org/>). DBPedia uses a namespace called the Simple Knowledge Organization System (SKOS; see <http://www.w3.org/TR/skos-primer/>) to introduce concepts such as “category.” Within DBPedia different solvents have been described as belonging to various categories, such as hydrocarbons or ethers. By combining the RDF statements from our data with that from DBPedia, it is possible to query our experimental data for examples of measurements done in different solvent categories. This is possible because the resource at RON links specific concepts (the identity of a molecule) to resources in both our data and DBPedia. This works

despite the fact that DBPedia might contain incorrect chemical names (as long as the resource links to the object with the correct InChI), that our data contains no concept of solvent category, and that DBPedia understands nothing about the ONS namespace.



**FIGURE 16-4.** Connecting solubility measurements with the wider data web via RDF. RON is <http://rdf.openmolecules.net>, the resource that connects records from DBPedia, Chemical Blogspace, and ChEBI (a European Bioinformatics Institute Chemistry resource).

Taking this one step further, we can link our experimental data into a wider discussion on the Web by using RDF from RON to identify, for instance, which blogs have been discussing a particular chemical compound. This RDF contains links to Chemical blogspace (<http://cb.openmolecules.net/>) and shares the use of unique identifiers (in this case, the InChI is used in URI form). The [rdf.openmolecules.net](http://rdf.openmolecules.net) resource links to a range of data sources, again providing a way for data and analysis from multiple sources to be combined together. The value of the RDF approach is that additional data sources can be added at any point on the graph, without having to worry about how that information relates to that from other data sources. Work can always go into making the integration better by choosing to share more common vocabulary elements, but as long as a new data source has at least one common identifier, then data integration can begin.

## Closing the Loop: Visualizations to Suggest New Experiments

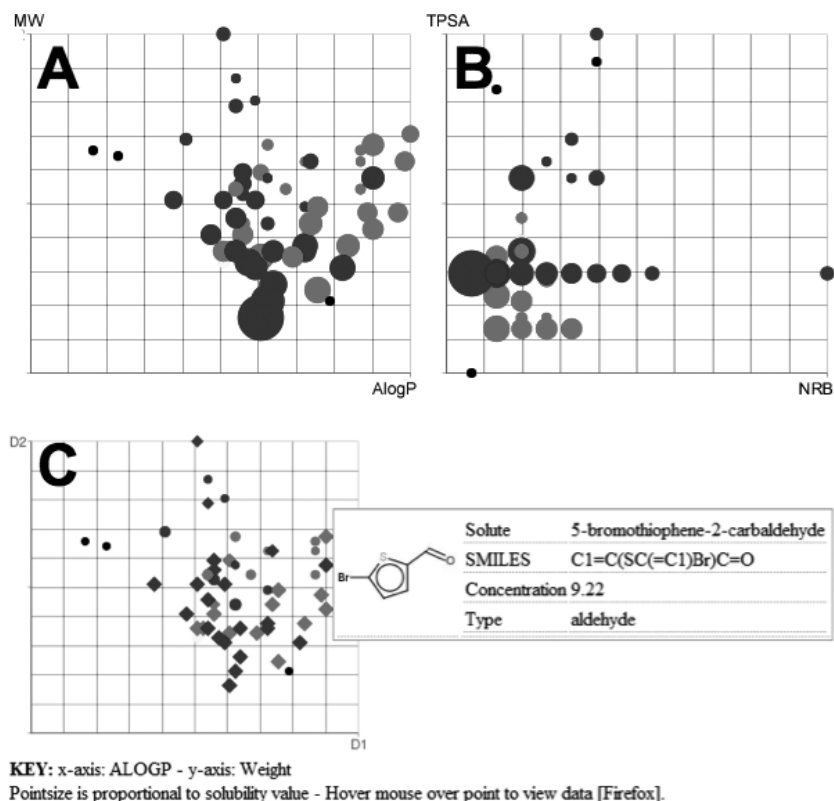
As noted earlier, data from experiments can be utilized in a variety of ways, ranging from visualization to modeling. These activities are useful and can provide insight into the physical problem at hand. However, our main aim is to use the modeling and analysis to inform the design of new experiments. As the crowdsourcing effort expands, it is important to consider possible experiments and prioritize these, particularly if the ultimate aim is to enable interested, but not necessarily experienced, researchers to take part. Such computational prioritization is very useful in many scenarios, where resources (financial,

material, time) are limited and all possible experiments cannot be carried out. In the case of solubility, an experimentalist might ask, "Given the compounds tested so far, which ones should we do next?" Visualization of the data can be both compelling and provide a good guide to the best choice of the next experiment given the resources available. This enables a cyclical relationship between experiment and computation, making optimal use of both the experimentalist's and analyst's skills.

To identify which compounds we, or anyone else, should test next, we need a way of understanding where in a "chemical space" each of the compounds we have already tested lie. Then it will be possible to identify empty parts of that space in our data set, correlate that with specific molecules that lie in those spaces, and carry out those experiments. This requires the integration of information that is not found in our data. We have the identity of compounds and solvent as well as the solubility, but we do not know the characteristics of the molecules, i.e., their position in our chemical space. To obtain this information, we need to create a "mashup" of chemical data using a variety of services. We have provided a simple REST-based interface to CDK (Steinbeck 2006) descriptors. A URL of the form <http://www.chembiogrid.org/cheminfo/rest/desc/descriptors/c1cccc1COCC> retrieves an XML document that contains multiple URLs, each one pointing to an XML document containing the value of the specific descriptor. The chemical space characteristics used in the visualizations shown here are the compound molecular weight (MW), predicted hydrophobicity (ALogP, a measure of the preference of the compound for water or oil solvents), and the calculated molecular surface area (TPSA). Many more descriptors are provided via the web service.

As all the services and data are provided on the open Web, it is possible for third parties to utilize these to prepare visualizations. Using the data in the GoogleDoc and the web services provided at Indiana, a visualization tool was independently developed that enables a multidimensional visualization of the solubilities of all compounds in a specific solvent (<http://oru.edu/ccda/sl/descriptorspace/ds.php>). In Figure 16-5, the X and Y axes display a specific molecular descriptor, the color identifies the type of compound, and the size or shape of each point shows the solubility. In addition, hovering over each point activates a tooltip giving further details, including structure and solubility. The figure shows clear areas of the chemical space that are not occupied by currently available data points (the bottom left of panel A, for instance). In principle, further services could be configured to suggest compounds that lie in those areas by querying data sources such as ChemSpider.

To expand the ability to display multiple dimensions, further visualizations were prepared in the 3-D environment of Second Life (<http://www.secondlife.com>; Figure 16-6). As with GoogleDocs, Second Life may seem an odd choice for a scientific visualization environment. However, once again it satisfies many of the criteria that we have applied to other parts of our project. It provides a (relatively) simple environment for the user in a generic package that is available free of charge. It therefore reduces barriers compared to other specially developed and often complex and expensive visualization environments. From the perspective of the visualization experience, Second Life also offers many advantages. It is possible to move around the graph, to zoom in and out, and even to walk inside and

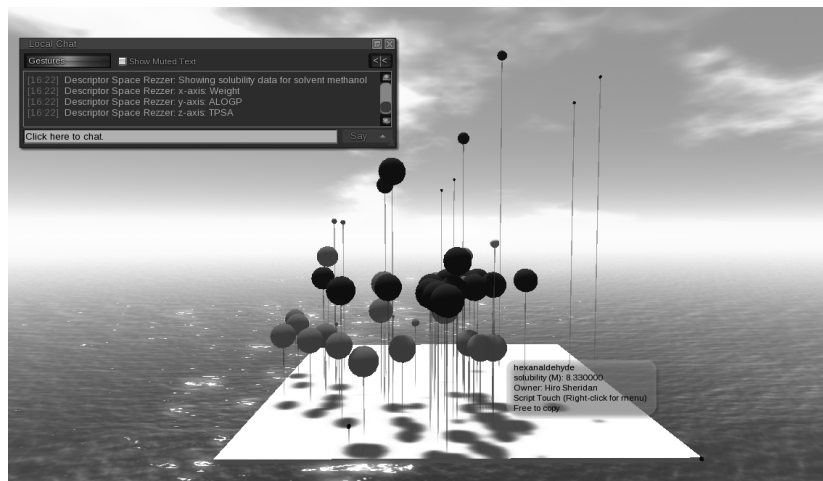


**FIGURE 16-5.** Graphical representation of solubility data in chemical space. Panels A and B show two visualizations of the same data plotted onto axes representing different chemical characteristics. The color of the spots represents the chemical type (red for aldehyde, blue for carboxylic acid, yellow for amine, and black for other) and the size the solubility. Panel C illustrates the clickable interface showing the chemical structure and value of the solubility for one data point. (See Color Plate 56.)

examine the graph from that perspective. Multiple users can also simultaneously view and manipulate the same graph. From the developer's perspective, Second Life provides the ability to bring in data from the outside Web, enabling the use of the web services described earlier, and provides a clickable interface for the user to manipulate the graph or to follow links from the data points to data source and experiments. In a theoretical sense it might be preferable to use an open source rendering system that worked entirely within a web browser, but open source systems are limited and no other system offers the combination of technical ability, simplicity of interface, and usability of Second Life. In a very real sense, these compelling visualizations speak for themselves.

It is clear that easy access to data allows computational scientists to perform a variety of analyses, but the close integration of experiment and computation allows the overall investigation to be much more efficient. Although many computational analyses require significant manual input and cannot be converted to an automated online service, many are simple enough to be converted to a service that can be incorporated into a variety of platforms. The result is that it is significantly easier to analyze and manipulate the data to





**FIGURE 16-6.** Representing multidimensional data using Second Life. Three chemical descriptors are represented on the three spatial axes. The color of the balls indicates the type of chemical entity (as defined in the previous figure), and the size shows the solubility in the current solvent. The visualization is available at <http://slurl.com/secondlife/Drexel/165/178/24> on Drexel Island, Second Life. (See Color Plate 57.)

suggest new directions for experiments, as well as develop novel applications, by mash-ing together data and applications. These mashups demonstrate the power of using well-recognized and easily convertible, machine-readable identifiers. The SMILES code in this case is the key identifier that can be used to obtain further data from other web services, data sources, or data from experiments exposed by other researchers. In the future, the use of RDF to describe the results has immense promise in allowing automated integration. As RDF provides a self-describing framework based on agreed dictionaries, it is possible to search for data services that provide the desired information without any prior knowledge of where they are or what their internal data schema is. Most current mashups work by using a single common key (geographical location, search term, date) on known services with a known schema. The promise of an open data web where the links between objects are self-describing is that anyone will have the ability to create arbitrary mashups in which the search for data and information sources is an integral part of the process.

## Building a Data Web from Open Data and Free Services

A large part of the art of performing and communicating science is in designing processes that remove inaccurate or misleading results, to provide a body of evidence that clearly supports a simplifying explanation that humans can understand. Science can be seen as the process of reducing pieces of the world into intelligible models. Part of the problem of this approach is the tendency to oversimplify to either strengthen an argument or, in the case of very complex systems, just to make it comprehensible.

Our approach is to embrace the complexity of real measurements by making all the detail available. We aim to balance the issues that this complexity creates with the need for clear and useful data sets by filtering the primary record in as transparent a way as possible to create the primary data set. The availability of storage space on the Web at near zero cost and the wide availability of high-quality, freely hosted services makes it possible to host the whole of the research record in the public view. There is simply no longer any excuse for writing “data not shown.” But the desire to provide access to the full record creates new problems.

The first of these is simply volume. The research record itself tends to be a large body of largely unstructured text and images. Widespread standards do not exist to represent this type of information in a way that is easily parsed by either humans or machines. Summaries and filtering are required to make available the information contained in the record. We chose to use a GoogleDocs spreadsheet as the primary source of extracted data. The process of extracting data from the record remains subjective and manual for the moment. The spreadsheet provides a natural interface for humans, and in particular experimental scientists, while also providing a range of effective interface options for web services to reprocess and represent the data.

It is possible to imagine scraping data directly from the experimental report. With a small amount of informal formatting and regular expression analysis and conversion of a feed generated from the record, it would be possible to automatically populate the spreadsheet. We have not pursued this, because we wish to include a human filtering process at this stage. As the project increases in size, this will become untenable at some scale. The choices about what scale will depend on the project, the type of data, and the need or desire for precision and accuracy in the presented data set.

Once the data is made public it is open for use by any interested researcher, and the Google-Doc API makes it possible to exploit the data for a wide range of services. This can include visualization or analysis services. These services will be dependent on understanding how the data is structured within the spreadsheet. This means they will generally be written against the specific data set. However, even in this case it is straightforward to leverage a wide range of services, data sources, and visualization tools to create highly effective data displays, ranging from tables and simple graphs to clickable interfaces in five, seven, or more dimensions. Open standards and open systems provide the ability to move data and information to the places where it can be most effectively used. The promise of truly open and self-describing data formats is immense but unrealized, even in data-driven sciences such as chemistry, due to both technical and social difficulties in translating from records in the form that experimentalists understand to properly structured machine-readable forms as understood by computers and the people who code on them. Here we have shown the ability to convert data in the form of a spreadsheet (something that experimentalists are familiar and comfortable with) to RDF, but other formats could be served just as easily.

The provision of such general data formats makes it possible to create services that integrate data from multiple sources. A wide range of data sources containing information on solubility, or perhaps other information in our data set, could be integrated and analyzed together. This will make it possible for aggregation and link farm services, such as Chem-Spider and others, not just to automatically aggregate data, something that is already technically feasible, but to make educated decisions about the level of curation required by data from different sources, and to deploy human curation where it is most needed. This centralization, in turn, provides a valuable indexing service, providing a central location online where users can find sources for the data they are looking for.

One of the central themes of all the work described here is the use of free hosted systems that provide enough functionality without overburdening the user with complexity. For most of the recording, aggregating, analysis, visualization, and presentation steps, there are more advanced, more general, or more sophisticated tools available. To record the research, we could have used a commercial Electronic Laboratory Notebook or a specially designed online system; instead, we used a freely hosted wiki service. The presentation of the primary data could have used a database backend with a content management system to provide sophisticated visualizations; we used an online spreadsheet and its API with some JavaScript to present a range of visualizations. There are many highly functional and sophisticated 3-D viewing environments available; we used Second Life.

Part of the rationale for this is cost. All the services we used are free at point of access, allowing an essentially unmanaged development process to grow and attract new collaborators with low barriers to their entry. However, a significant part of the rationale is to use services that are fit for purpose without being overly sophisticated. Forging an effective link between experimentalists and analytical and theoretical scientists is always a challenge. The use of the spreadsheet as a source of data that can be automatically or manually converted to well-described formats (a formal relational database), or self-describing and extensible formats (RDF), or simply readily transformed directly into sophisticated visualizations, illustrates the use of the spreadsheet as a meeting point. Experimental scientists like and understand spreadsheets. Computational scientists may prefer either text files over which code can be run, databases, or formats such as XML and RDF. Key to bringing these communities together will be the ability to convert backward and forward into preferred formats in a fully automated fashion.

Finally, the key to the whole project is trust and transparency. As a record is converted into data, and data is converted into information, and finally as information is converted into a model or theory, context is lost at every stage. The details, which are often messy, get left behind so that the bigger picture can emerge. This is entirely appropriate. Science is the process of summarizing observations in a way that allows us to predict what will happen in the future. As with our choice of services, a scientific model or theory is useful if it is good enough to save us the time of doing the experiments most of the time. Traditionally, however, this summarizing process has come at the price of losing access to the detail. In the world of the Web, where storage is cheap, this no longer needs to be the

case. Now the choice lies in how to present the underlying detail, what filters to apply in the summarizing process, and how to retain the links between the summaries and the original records.

These are not easy decisions, and we would not claim to have got all of them right. Nonetheless, we believe this project can act as an exemplar of the approach. Over the course of four months, a project that started as a discussion between two people on a train has grown into a multinational data collection, visualization, and modeling effort where all participants have access to all of the data and analysis in real time. The collaboration can easily grow as new researchers become interested. Our open data and open services have enabled the creation of compelling new visualizations without requiring any direct involvement from the experimentalists themselves. These visualizations are both useful to the experimentalists and strikingly beautiful in their own right. Yet they represent only a small proportion of what could be done, by anyone, with the data we have exposed. At the same time they always provide a link back to the original record, with all its blemishes and weaknesses, allowing any user to assess the validity and strength of any specific data point at the resolution she chooses.

Beauty is often seen as being related to simplicity or symmetry, a sense that the whole can be described using a simple mathematical description. This is rarely the case with real experimental data. The beauty that lies, sometimes hidden, within experimental data may take extensive filtering to reveal. But, if the true beauty lies in understanding, as far as we can, what is really happening at the deepest possible level in the world around us, then we can uncover only a limited amount of that beauty through any given analysis. By providing access to as much of the record as we can, we make it possible for other researchers to discover and reveal more of the beauty that is hidden beneath the surface.

## Acknowledgments

The authors would like to acknowledge the efforts of the researchers who collected the majority of the data, Khalid Mirza, Jennifer Hale, and Tim Bohinski; Bill Hooker for assisting with judging of the Open Notebook Science Challenge; and the financial and in-kind support of Submeta and Nature Publishing Group.

## References

Bradley, Jean-Claude. 2007. "Open Notebook Science Using Blogs and Wikis". Available from *Nature Precedings*, <http://dx.doi.org/10.1038/npre.2007.39.1>.

Bradley, Jean-Claude et al. 2008. "Optimization of the Ugi reaction using parallel synthesis and automated liquid handling." *Journal of Visualized Experiments*, <http://dx.doi.org/10.3791/942>.

Steinbeck, C. et al. 2006. *Current Pharmaceutical Design*, 12, 2110–2120.

