# CDI-Type I: Chemistry Crowdsourcing using Open Notebook Science

*PI: Jean-Claude Bradley, Drexel University*

## Abstract

The current system of dissemination of scientific data and knowledge is far less efficient than it needs to be to facilitate improved collaborative science, especially considering current publication vehicles and infrastructure. There is a growing movement promoting more Open Science with the belief that a more transparent scientific process can perform far more effectively. The logical extension of this concept is full transparency - exposing a researcher's complete record of progress to the public in near real time. Not only will such a process enable ongoing data sharing it also provides an opportunity to develop collaborative communities of scientists and, at the conclusion of data acquisition, can enable communal extraction of conclusions when necessary. We have named this approach Open Notebook Science and have demonstrated its implementation and feasibility with the UsefulChem project, started in the summer of 2005, with the aim of synthesizing novel anti-malarial compounds. Our system currently uses free hosted services using general blog and wiki functions to facilitate replication across any scientific domains. These services are not chemically intelligent and are limited to text and graphic based data sharing only. For Open Notebook Chemistry the ability to intelligently manipulate, manage and search chemical structures and associated data is necessary and we have demonstrated proof of concept capabilities by integrating with the ChemSpider service, a free access online database managing chemical structures and focused on developing a structure centric community for chemists. This work will require the development of a chemically intelligent software platform to extend the capabilities of both the blog and the wiki environment for managing Open Notebook Science. The exposure of raw experimental procedures and data in a semantically rich format will enable the participation of both human and autonomous agents in the process of scientific discovery. This phenomenon of spontaneous group intelligence, referred to as "Crowdsourcing", has proven valuable in several contexts. Already, productive collaborations have been forged within the UsefulChem project with groups from Indiana University, Nanyang Technological University, the National Cancer Institute and UC San Francisco.

Nature Precedings : doi:10.1038/npre.2008.1505.1 : Posted 9 Jan 2008

## Intellectual Merit

The trends of Open Science, crowdsourcing, automation and cheminformatics are creating new opportunities for increasing the efficiency of discovery. The integration of these phenomena promises to enable new forms of scientific collaboration. This project brings together people who are uniquely qualified to create an infrastructure to realize that potential.

## Broader Impact

By its very nature, crowdsourcing is designed to connect people together in constructive and unexpected ways. A key component of this project is to leverage the knowledge gained from an experiment in crowdsourcing in a particular field and make it easier for it to be applied in other scientific collaborations. The open nature of the project ensures that this knowledge will be disseminated nearly in real-time and others interested in replicating such systems will have immediate access to advice and resources from the participants. Already this replication effect has taken place, with other laboratories exposing their work to Open Notebook Science conditions, stimulated by reports of the UsefulChem project.

## *List of Participants*

### Jean-Claude Bradley, Drexel University, PI:

Jean-Claude Bradley is an Associate Professor of Chemistry and E-Learning Coordinator for the College of Arts and Sciences at Drexel University.  He leads the UsefulChem project, an initiative started in the summer of 2005 to make the scientific process as transparent as possible by publishing all research work in real time to a collection of public blogs, wikis and other web pages.  Jean-Claude coined the term Open Notebook Science to distinguish this approach from other more restricted forms of Open Science.  The main chemistry objective of the UsefulChem project is currently the synthesis and testing of novel anti-malarial agents.  The cheminformatics component aims to interface as much of the research work as possible with autonomous agents to automate the scientific process in novel ways.  Jean-Claude teaches undergraduate organic chemistry courses with most content freely available on public blogs, wikis, games and audio and video podcasts.  Openness in research meshes well with openness in teaching.  Real data from the laboratory can be used in assignments to practice concepts learned in class.  Jean-Claude has a Ph.D. in organic chemistry and has published articles and obtained patents in the areas of synthetic and mechanistic chemistry, gene therapy, nanotechnology and scientific knowledge management.

### Kevin Owens, Drexel University, co-PI

Kevin Owens is an Associate Professor of Chemistry at Drexel University. His research group focuses on the use of mass spectrometry for both biological and synthetic polymer analysis, specifically mechanistic studies of the Matrix-Assisted Laser Desorption/Ionization (MALDI) process, application of MALDI to synthetic polymer analysis, development of sample preparation methodologies for quantitative analysis by MALDI Time-of-Flight Mass Spectrometry (TOFMS), TOFMS instrument development, and the application of chemometric techniques (particularly correlation analysis) to aid in the automated interpretation of mass spectral data. His latest work in the area of mass spectral interpretation involves the combination of genetic algorithm techniques with correlation analysis to create mass spectral filters for the automated identification of chemical substructures. His group maintains a number of wikis describing several of their current research efforts. Kevin has a Ph.D. in analytical chemistry, and teaches the two-quarter undergraduate instrumental analysis sequence, the graduate/undergraduate chemical information retrieval course, as well as the graduate level mass spectrometry and statistics & experimental design courses.

### Antony Williams, ChemSpider, Senior Personnel

Antony Williams is a Senior Fellow at the National Institute of Statistical Sciences, is the President of ChemZoo Inc., and is the host of ChemSpider, a Structure Centric Community for Chemists.  Antony has worked in academia, in industry (Eastman-Kodak) and in the commercial cheminformatics sector in the role of Chief Science Officer.  He has authored and co-authored over 100 scientific publications and multiple invited book chapters. He has two issued patents and has one pending. His formal training is as an NMR Spectroscopist with a focus on small molecule structure elucidation and analytical data processing algorithms. the last decade of his career has been focused on the development of integrated sample, structure and analytical data management systems at the desktop and at the enterprise level utilizing web-based technologies. In the past year he has established a free access website, ChemSpider, to provide access to chemistry-related information for almost 20 million chemical entities and their associated data. He is interested in all aspects of cheminformatics, with special interest in chemical structure handling, nomenclature and computer-assisted structure elucidation. He conducts research in the area of data processing techniques for spectroscopy and development of the semantic web.

## *Justification of Intellectual Partnership*

This project aims to synergize the following tools and processes:

**1) Crowdsourcing**

**2) Open Notebook Science**

**3) Automation and reaction optimization**

**4) Cheminformatics**

Crowdsourcing is a term introduced by Jeff Howe to describe the solution of problems through a distributed network of people.(1) Although there are several examples of crowdsourcing in science, most are not open. The best known example is Innocentive, where scientific problems are made public with prizes for the solvers.(2) However, the proposed solutions are not made public and accepted solutions are intended to be proprietary to the company funding the solution. Innocentive is also collaborating with the Rockefeller Institute for help to combat third world and orphan diseases.(3) Other non-commercial examples in science include Stadust@home (4) and GalaxyZoo (5), initiatives designed to identify astronomical objects. Some recent examples of crowdsourcing in chemistry are Chemmunity (6), the Synaptic Leap (7), OrgList (8), Chemists Without Borders (9) and ChemUnPub (10).

The term "Open Notebook Science" was coined to represent a form of Open Science where the laboratory notebook is made public in as close to real time as possible.(11) The Bradley lab has demonstrated the feasibility of carrying out Open Notebook Science since the summer of 2005 with the UsefulChem project. Experiments are stored on wiki pages, much like in a paper notebook, but with hyperlinks to all the raw data collected.(12) There is also a blog to report on the overall progress of the scientific work. (13) The blog and wiki receive about 200 hits per day. Over time, a collaboration with other scientists has evolved. Rajarshi Guha at Indiana University and Tsu-Soo Tan from Nanyang Polytechnic in Singapore have invested significant amounts of time in running docking calculations for UsefulChem virtual libraries and reporting their results openly, in near real-time. Dan Zaharevitz, from the National Cancer Institute has contributed by testing compounds for potential anti-tumor activity. Phil Rosenthal, from UCSF, is in the process of testing some compounds for anti-malarial activity.(14) Other individuals from around the world have contributed by engaging in open discussions in the blogosphere. The UsefulChem project has also been cited in the peer-reviewed literature. (15, 16)

These examples of spontaneous collaboration clearly indicate that there is a huge potential for doing science under open conditions. However, in order to take full advantage of this potential, additional elements must come into play. Currently, experimental results on the wiki pages are written in a format amenable to human interpretation. The system could be made much more powerful by representing the information in a format understandable by machines as well, so as to contribute to the growing semantic web.

As a step in this direction, each experiment page has a “tag” section at the end where more machine-friendly representations of molecules used in the experiment are listed. As an example, InChI codes are alphanumeric representations of molecules using standards supported by IUPAC. Since these InChIs are indexed by search engines, such as Google, it permits an unambiguous means of retrieving relevant information about each chemical. Other means of representing chemicals (SMILES, chemical names) suffer from having non-unique representations.

This tagging system works for simple retrieval of an exact chemical structure but fails to allow more sophisticated queries, such as, identifying chemical substructures, similarity of structure, whether a chemical is a reagent or a product, the reaction conditions, and details regarding characterization of the materials, etc. ChemSpider (17) is a free access online database working to build a structure community for chemists and offers some of the necessary capabilities to facilitate this crowdsourcing project. The system has been built specifically with the intention of hosting chemical structures and related information

Nature Precedings : doi:10.1038/npre.2008.1505.1 : Posted 9 Jan 2008

such as analytical data. The system presently hosts almost 20 million chemical structures and is the first online system to facilitate public depositions of single structures or collections, property data, unstructured annotations and analytical data. ChemSpider (CS) has been designed with the needs of the chemist in mind and over 3000 chemists per day use the system for the purpose of searching for data and information associated with chemical entities. The system has already started to deliver on its promise to facilitate connectivities via the semantic web by providing open web services to allow other platforms to integrate and derive value from the information available online. CS has already committed to two directions of related interest to this project : Development of a platform for collaboration between chemists (18) and the development of a structure-based encyclopedia for chemists, WiChempedia (19). These developments will provide core functionality to support our Open Notebook Science efforts. Additional capabilities will be introduced on an as-needed basis to facilitate our efforts in terms of the support of experimental data capture, control and reporting as outlined below.

The final component of this partnership is the automation of the execution of some of the experiments. This will allow us to take scientific crowdsourcing to the next level, where the online community can not only provide feedback but also directly carry out simple experiments.  We propose to incorporate a ChemSpeed Technologies AG Accelerator SLT100 Synthesizer system into the proposed workflow. (20) The Accelerator is a fully-automated robotic synthesis platform capable of both solution and solid phase synthesis. The unit holds a maximum of 12 reactor arrays, where each of the arrays contains a number of identical reactors with available volumes between 2-100 mL. For example, there are 16 reactors/array using 2 mL or 13 mL reactors. The reactors can be heated and cooled over the range of -70 to +180C and they can be vortexed up to 1400 rpm. An inert or reactive gas blanket or vacuum (down to 10 mbar) can be applied to the array for evaporation. The robotic system is equipped with a pipetting head composed of 4 independent needles attached to 2-way 6-port ceramic valves enabling the direct use of 16 solvents. The unit has solid dispensing capability with 1mg-10g capacity (with a resolution of 0.1mg). Each reactor array may be equipped with automated filter capability with either the filtrate or precipitate available for subsequent reaction steps. An HPLC injection valve or other analytical capabilities are available as options. The ChemSpeed Autosuite instrument operation software includes a simple drag and drop programming interface built on an SQL database. We expect this will allow easy integration with the open source concept which serves as the basis of this proposal.

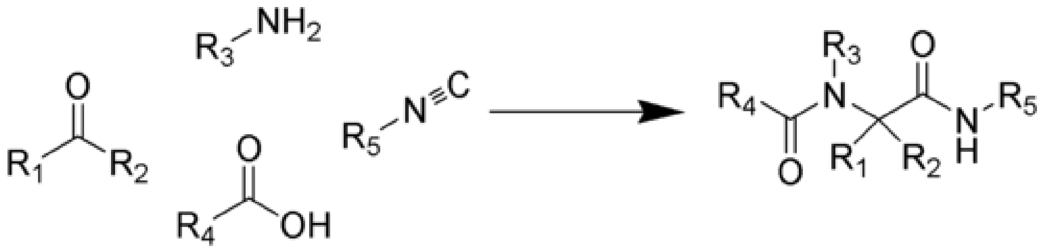
## *Research program*

The following scientific objectives will be initiated as projects:

### 1) Determining optimal conditions for the precipitation of Ugi products.

The Ugi reaction provides a rapid access to large combinatorial spaces by bringing together an amine, an aldehyde, a carboxylic acid and an isonitrile.(21)  In addition, reaction conditions are generally very convenient: methanol at room temperature is a common protocol.  The Bradley group has used the Ugi reaction to synthesize potential anti-malarial compounds.  They observed that sometimes the products simply crystallized from the reaction mixture.(22)  Although NMR monitoring indicated that the reaction proceeded smoothly in most cases, the products did not always precipitate.  The ability to purify products by a simple filtering step translates into a significant advantage compared against the alternatives, which would usually consist of chromatography, a costly and inefficient process that severely limits scaling up.

Nature Precedings : doi:10.1038/npre.2008.1505.1 : Posted 9 Jan 2008



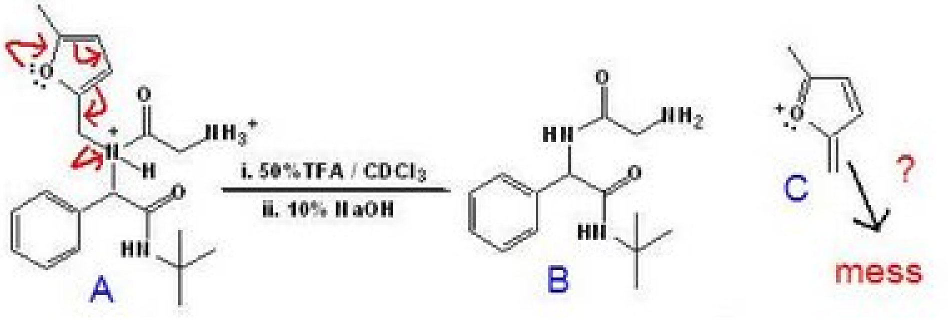
Since the Ugi reaction is of value to the generation of product libraries in multiple applications, the ability to predict the crystallization event should be of significant value to the chemistry community. By generating models to accurately predict precipitation, virtual libraries being screened for a desired activity can be further screened for ease of preparation. A better understanding of the solubility of Ugi products in different solvents using more sophisticated models may also recommend different reaction conditions for those products that do not crystallize from the standard protocol using methanol.

The simplicity of this initial project is well suited for crowdsourcing. The results (precipitate/no precipitate) are simple and easy to tabulate. Anyone familiar with QSAR analysis can propose a model and suggest combinations of starting materials to most efficiently test it. The experiments are simple to execute either manually or with the assistance of automation. In the case of precipitates, all products will be characterized by standard techniques (H NMR, C NMR, IR, MS, HRMS).

In addition to the general crowdsourcing effort, co-PI Kevin Owens will leverage his knowledge of experimental (i.e., factorial) design and optimization methodology (e.g., simplex optimization) to help guide synthetic reaction optimization. With respect to the prediction of precipitation, the genetic algorithm approach can quickly build a diverse set of Ugi products created using the parallel synthesis capability of the automated chemical reactor system requested as part of this proposal. The wide compound diversity will allow the QSAR methods to more quickly explore the experiment space and develop a robust predictive capability.

## 2) Mechanistic investigations

This project will also explore the effectiveness of crowdsourcing to solve mechanistic problems. An initial problem will focus on the acid-catalyzed cleavage of furfuryl groups.(23) It has been observed that compounds such as A cleave to yield de-furfurylated product B and some intractable material, perhaps resulting from the polymerization of side product C. A mechanism proposed to account for this behavior is shown below. Derivatives of A will be synthesized and the rate of this transformation measured via NMR spectroscopy. Electron donating groups on the furfuryl ring are predicted to accelerate the cleavage while electron-withdrawing should slow it. Quantitative models for this reaction, specific test reactions, and alternative mechanisms will be sought from the online chemical community.



### 3) Other projects will be selected from crowdsourcing efforts

Over time, suggestions will be taken from the online community for additional problems of importance that can be solved with our growing infrastructure of students, collaborators and informatics and automation tools.

## *Educational program*

This project offers educational opportunities for both graduate and undergraduate students at several levels. The following skills will be acquired:

### 1) Organic and analytical chemistry skills

Students working in the Bradley laboratory will learn standard organic chemistry techniques, including executing experiments then isolating and fully characterizing products with conventional spectroscopic analysis (NMR, IR, MS, HRMS). Students working in the Owens lab will learn analytical skills, including qualitative and quantitative mass spectrometry techniques. Students in both groups will learn and apply the principles of experimental design and optimization for improved synthesis and analysis..

Nature Precedings : doi:10.1038/npre.2008.1505.1 : Posted 9 Jan 2008

### 2) Networking

Students will be required to record their experiments using an Open Notebook on a wiki and will receive and respond to comments from the online world. Maintaining a public laboratory notebook can be a very efficient way to learn about the proper way to document an experiment because the adviser and other interested parties can provide immediate and ongoing feedback, which is impossible with a conventional closed paper notebook. They will also be encouraged to engage in conversation about their project on various social networks, including mailing lists (e.g. OrgList, UsefulChem), our collaborators' blogs and wikis, Facebook, Nature Networks, SciVee, Flickr, etc. Not only will interacting with peers and mentors be valuable as a learning experience, the contacts formed may be helpful for the progress of the student's career after graduation.

### 3) Automatic reactor programming

Students will be charged with setting up, running and maintaining the automated chemical synthesis system. Coupled with their work doing conventional manual organic synthesis, these students will be competitively trained to enter the 21st century chemistry workforce.

### 4) An understanding of cheminformatics

In order for the experiments to be properly indexed by ChemSpider and other automatic aggregators, students will have to learn how to generate and use representations of molecules using modern cheminformatics techniques (e.g. use structure drawing tools for the generation of SMILES, InChI, and InChIKey molecular representations). They will also need to become familiar with tools for the manipulation of analytical data, image management tools, etc. and will learn to order their data management processes and behaviors in a logical manner in order to capture, document and mine data in an electronic environment. It is becoming increasingly important for chemists to be fluent with these representations to optimally use software and online databases.

## Timeline

### Year 1:

During the first year, the main focus will be on finding collaborators from the online community. Experiments will be performed manually until the automatic reactor has been installed and students trained to operate it. Informatics tools to represent experimental workflows and reactions through their various entity-state-increment-transition relationships will be developed. In a reaction workflow at any point in time chemical entities are involved and are at a particular state (mixed, unmixed, heated, stirred) for a particular time increment and then pass through a transition to another state (analytical data acquired, new material added etc.). These relationships can be described in a logical sequence and described in a manner allowing the data to be mined throughout a workflow sequence. Informatics tools to describe, capture, manage and datamine will be developed.

### Year 2:

The automated reactor will be implemented and made available for direct operation by the crowdsourcing community under the oversight of the PI and co-PI. Enhanced workflow management tools to generate workflow documents capable of driving the automation systems will be developed. Data will be generated in a standard manner to allow data capture and management on the ChemSpider platform.

### Year 3:

Projects suggested by the crowdsourcing community will be evaluated and resources will be allocated to execute the experiments, under supervision of the PI and co-PI. As sub-projects get completed, work will be submitted for traditional peer reviewed publication. Arguments in the papers will be supported by links to the original experiments in the online open notebook wiki, giving unprecedented systematic access to experimental raw data to be re-analyzed or re-purposed by anyone. Co-authorships will be based on documented contributions from members in the community who sufficiently participated.

## Desired Project Outcomes

1) From the larger perspective a key outcome is the establishment of a precedent and model to demonstrate how crowdsourcing can be used to solve scientific problems. As the project evolves progress will be reported on social software networks. This type of dissemination has proved effective to identify collaborators in the online world for the UsefulChem project, primarily through the use of blogs. Another key advantage is that reports of what appears to work or not work are generally followed quickly with helpful discussions in the blogosphere, in addition to being helpful to others with related projects. The demonstration that an automatic reactor can be effectively added to the scientific crowdsourcing toolkit is also a key outcome.

2) At a more basic level, meeting the scientific objectives outlined in this proposal will benefit the scientific community in a more direct way. The Ugi reaction is already used by many groups and further understanding of how to prepare these products using a much simpler purification protocol can be quite valuable in many areas. For example, several of the Ugi products from UsefulChem project have been predicted to possibly act as anti-malarial and anti-tumor agents and testing is underway.

## References

1) Howe, J. The Rise of Crowdsourcing, Wired, June 2006.
(http://www.wired.com/wired/archive/14.06/crowds.html )

2) http://www.innocentive.com

3) http://www.innocentive.com/servlets/project/Pavilion.po?p=Rockefeller%20Foundation

4) http://stardustathome.ssl.berkeley.edu

5) http://www.galaxyzoo.org

6) http://www.chemmunity.com

7) http://www.thesynapticleap.org

8) http://www.orglist.net

9) http://www.chemistswithoutborders.org

10) http://www.chemunpub.it

11) http://drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html

12) http://usefulchem.wikispaces.com/All+Reactions

13) http://usefulchem.blogspot.com

14) http://usefulchem.blogspot.com/2007/12/first-falcipain-2-targets-shipped.html

15) Todd, M. Open Access and Open Source in Chemistry, Chemistry Central Journal 2007, 1:3.
(doi:10.1186/1752-153X-1-3)

16) Lancashire, R. The JSpecView Project: an Open Source Java viewer and converter for JCAMP-DX, and XML spectral data files, Chemistry Central Journal 2007, **1:**31. (doi:10.1186/1752-153X-1-31)

17) http://www.chemspider.com

18) http://www.chemspider.com/blog/the-chemspider-team-chooses-our-future-platform-for-collaboration-microsoft-sharepoint.html

19) http://www.chemspider.com/blog/wichempedia-is-now-on-its-way.html

20) Chemspeed, Augst, Switzerland http://www.chemspeed.com/index.php

21) Ugi, I.; Werner, B.; Dömling, A. (2003). "The Chemistry of Isocyanides, their MultiComponent Reactions and their Libraries". *Molecules* **8**: 53-66.

22) http://usefulchem.blogspot.com/2007/11/combiugi-update-master-table.html

23) http://usefulchem.blogspot.com/2007/05/missing-methyl-mystery-mechanism.html