The Quixote project: Collaborative and open Quantum Chemistry data management in the Internet age

Pablo Echenique[1,2,3] , Jorge Estrada*[1,2,4] , Peter Murray-Rust , Jens Thomas , Please add the rest of you

[1]Instituto de Química Física "Rocasolano", CSIC, Serrano 119, E-28006 Madrid, Spain
[2]Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Mariano Esquillor s/n, Edificio I+D, E-50018 Zaragoza, Spain
[3]Departamento de Física Teórica, Universidad de Zaragoza, Pedro Cerbuna 12, E-50009 Zaragoza, Spain
[4]Departamento de Bioquímica y Biología Molecular y Celular, Universidad de Zaragoza, Pedro Cerbuna 12, E-50009 Zaragoza, Spain

Email: Pablo Echenique: echenique.p@gmail.com; Jorge Estrada*: jorge.estrada@unizar.es;

*Corresponding author

Abstract

Background: Despite the increasing accuracy and cost of high-quality quantum chemical calculations, the valuable data produced never leaves, in most cases, the hard disks of the groups that calculated it, and even there the information becomes gradually unreadable or unlocatable. In contrast with other disciplines like crystallography, or bioinformatics, where standard formats and well-known, unified databases exist, the situation in computational Quantum Chemistry is clearly suboptimal, thus decreasing the efficiency of the field, and related ones.

Results: In the Quixote project, we have developed the basic infrastructure to parse and convert quantum chemical data to semantically rich, standard markup formats; automatically act on a large number of files and perform complex operations on their directory structure, as well as batch conversions; upload the new machine-readable files, together with the original ones to open online servers; and expose the resulting database via a RDF triplestore that can be queried using SPARQL. We have done this in a few of months, with no extra funding and no centralized organization, only with the manpower and skills of a number of motivated researchers and the use of collaboratory technologies such as wikis, software repositories and VOIP multiconferences. **(we should probably add some things here, such as the dictionaries, and maybe say some things differently after we have finished the rest of the paper)**

Conclusions: We show how a difficult scientific, technical and social problem which has remain unsolved for decades, such as the management of data in Quantum Chemistry, can be successfully tackled in a short amount of time by a non-hierarchical group of researchers distributed in several countries. Not only will these results help to change the data management model in the field, but the methodology can be applied to other pressing problems related to data in computational and experimental science.

Background

Recently, high-level quantum chemical (QC) methods have become available to the broader scientific community through a number of user friendly software packages such as Gaussian [1], GAMESS-US [2], NWChem [3], MOLCAS [4] and many more. Additionally, the cost of computer power has experienced an exponential reduction in recent decades and, more importantly, sophisticated approximations have been developed that pursue (and promisingly approach) the holy grail of linear scaling methods [5,6]. This has enabled any researcher, with no specific QC training, to perform calculations on large, interesting systems using very accurate methods, thus generating a large amount of valuable and expensive data. Despite the scientific interest of this data and its potential utility to other groups, its lack of homogeneity, organization and accessibility has been recognized as a significant problem by important agents within the scientific community [7,8].

These problems, and specially the ones related to the accessibility of data have many consequences that reduce the efficiency of the field. As mentioned, QC methods are computationally expensive: The scaling of the computer effort and storage of high-level computations with the size of the system (N) is harsh, reaching, for example, N^7, for the most expensive and most accurate wavefunction-based methods, such as Coupled Cluster [9–11]. This makes it very difficult for groups that cannot use supercomputing facilities to have access to high-quality results, even if they possess the expertise to analyze and use the data. Even groups that do have access to powerful computational resources, given the lack of access to previously computed data by other researchers, often face the choice between *two inefficient* options: either they spend a lot of human time digging in the literature and contacting colleagues to find out what has already been calculated, or they spend a lot of computer effort (and also human time) calculating the needed data

themselves, with the risk of needlessly duplicating work.

Another problem originating in the lack of access to computed QC data and the very large number of methods available, is that users typically do not have the integrated information about which method presents the best accuracy vs. cost relation for a given application. The reason is that comparing one quantum chemical method with another, with classical force fields or with experimental data is non-trivial, the answer frequently depending on the studied molecular system and on the physical observable sought. Moreover, all the details and parameters that define what John Pople termed a *model chemistry* [12], i.e., the exact set of rules needed to perform a given calculation, have different and often opposite effects on the accuracy of the measured quantities in complex cases. As a consequence, the quality of the results does not steadily grow with the computational effort invested, but rather there exist certain tradeoffs that render the relation between them more involved [13–15]. Hence, not only the choice of the more efficient QC method for a given problem among the already existing ones, but also the design of novel model chemistries becomes 'more an art than a science' [16], based more on know-how and empiricism than in a set of systematic procedures.

These issues, and undoubtedly more that will appear in the future, together with a wealth of scientific problems in neighbouring fields, could be alleviated by public, comprehensive, up-to-date, organized, on-line repositories of computational QC data. Such infrastructures would increase the efficiency of the field, as it has been the case in crystallography **(please fill here details and references)** or in more experimental areas, like genetics or proteomics, where the NCBI GenBank[1] or the Protein Data Bank [2] constitute very successful examples of data sharing and organization. In an age in which both the monetary cost and the accuracy of QC calculations rival those of experimental studies, the need to extrapolate the model to this field seems obvious. The present situation, however, is far from the desired one.

On the one hand, there exist some in-house solutions that individual research groups or firms have built in order to implement a local-scale data management solution: This is the case of David Feller's Computational Results Database[3] [17], an intra-lab database to store and organize more than 100,000 calculations on small to medium-sized molecules, with an emphasis on very high levels of the theory. Also, the commercial standalone application SEURAT[4] can open and parse QC data files and allows for metadata customization by the user, thus providing some limited, local databasing capabilities. In the

[1] http://www.ncbi.nlm.nih.gov/genbank/
[2] http://www.rcsb.org/pdb/home/home.do
[3] http://tyr3.chem.wsu.edu/~feller/Site/Database.html
[4] http://www.synapticscience.com/seurat/

same family of solutions, ChemDataBase [18] is a data management infrastructure mainly focused in virtual screening which presents the distinctive feature of being able to create and retrieve databases over grid infrastructures. Packages for interacting with QC codes (launching, retrieving and analyzing calculations), such as ECCE[5] or Ampac[6], have modest data management capabilities too, although only insofar it helps to perform their main tasks, and they can be regarded as intra-lab solutions as well. Probably the most complete in-house infrastructure of which we are aware of is the RC^3 (Regional Computational Chemistry Collaboratory) developed by the group of David Dixon at the Department of Chemistry of the University of Alabama. The main objective of RC^3 is to perform the everyday data backup, collection and metadata assignment for calculations, and to organize them for research purposes. At the time of writing, RC^3 has been tested by 36 users for more than a year, and backed-up and organized 1.6 million files, amounting to 1.5TB of data storage. The database contains 144,000 records and it can currently parse multiple QC data formats **(please fill here details and references for RC3 but also for the rest of solutions mentioned in this paragraph; the intention is being brief though)**.
A different category of data management solutions from the one discussed above is that constituted by a number of online web-based repositories of QC calculations, normally developed by one research group with a very specific scientific objective in mind. Among them, we can mention the NIST Computational Chemistry Comparison and Benchmark DataBase (CCCDB)[7], which contains a collection of experimental and calculated ab initio thermochemical, vibrational, geometric and electrostatic data for a set of gas-phase atoms and small molecules; the Benchmark Energy and Geometry DataBase (BEGDB)[8] [19], which includes geometry and energy CCSD(T)/CBS calculations as well as other high-level calculations, with a special emphasis on intermolecular interactions; the DFT Database for RNA Catalysis (QCRNA)[9] [20], which contains high-level density-functional electronic structure calculations of molecules, complexes and reactions relevant to RNA catalysis; the Atomic Reference Data for Electronic Structure Calculations[10] [21] compiled at NIST, containing total energies and orbital eigenvalues for the atoms hydrogen through uranium, as computed in several standard variants of density-functional theory; or the thermochemistry database at the Computational Modeling Group of Cambridge's Department of Chemical Engineering[11], collecting thermochemical data of small molecules, powered by RDF and SPARQL and

[5] http://ecce.emsl.pnl.gov/index.shtml
[6] http://www.semichem.com/ampac/afeatures.php
[7] http://cccbdb.nist.gov/
[8] http://www.begdb.com/
[9] http://theory.chem.umn.edu/QCRNA/
[10] http://www.nist.gov/pml/data/dftdata/index.cfm
[11] http://www.nist.gov/pml/data/dftdata/index.cfm

offering the output files of the calculations, together with the parsed CML[12] [22] **(please fill here details for the solutions mentioned in this paragraph; the intention is being brief though)**.

Apart from these all-encompassing solutions (either local or web-based), in which one or a few groups build a complete data management infrastructure, one can also consider the possibility of adopting a modular approach, in which different researchers tackle different parts of the problem, whilst always enforcing the maximum possible interoperability between the modules. The BlueObelisk group (http://blueobelisk.sourceforge.net/wiki/Main_Page) [23] has been championing this approach for a number of years now, and many of the developers of the tools discussed below are members of it. In this category of solutions, we can also mention the Basis Set Exchange (BSE)[13] [17, 24], which provides an exhaustive list and definition of the most common basis sets used in QC calculations, thus facilitating the definition and implementation of semantic content regarding the method used, as well as improving the interoperability among codes at the level of the input data; modern tagging and markup technologies like XML and RDF together with the building of semantic dictionaries, not only to promote interoperability, but to do it in a web-friendly manner that allows one to easily plug modules and build complex online data management projects; the CML language (a chemical extension of XML)[14] [22] is also one of the few cases in which a common semantics has been widely adopted by the chemistry community, and its extension to the QC field is one of the cornerstones of the Quixote project described here. Also on the interoperability front, we can mention the cclib[15] [25] and CDK[16] [26] libraries, as well as the OpenBabel toolbox[17], which provide many capabilities for reading, converting and displaying QC data in many formats. Regarding the ease of use of possible data management solutions, the open source molecular editor and visualizer Avogadro[18] can certainly be used as a useful module in complex projects, and in fact the design of Quixote is being carried out in collaboration the developers of Avogadro, with the intention of efficiently interfacing it in future versions. The Java-based viewer Jmol[19] could perform similar tasks **(please fill here details for the solutions mentioned in this paragraph; the intention is being brief though)**.

All in all, and despite the numerous efforts described above, it is clear that a global, unified, powerful solution to the management of data in QC does not exist at present; at the same time that the new

[12] http://cml.sourceforge.net
[13] https://bse.pnl.gov/bse/portal
[14] http://cml.sourceforge.net
[15] http://cclib.sf.net
[16] http://cdk.sf.net
[17] http://openbabel.org
[18] http://avogadro.openmolecules.net
[19] http://jmol.sourceforge.net/

internet-based technologies, the existence of vibrant communities, and the wide availability of powerful software to perform the calculations, and to convert and analyze the results, all seem to indicate that the field is ripe to produce a revolutionary (and much needed) change in the model. In this article, we present the beginnings of an attempt to do so.

Results and Discussion

(After a brief paragraph stating that we describe the methodology in Methods, we say that the use cases are our results and then we describe them)

Use case 1

Use case 2

Conclusions

Each day, countless calculations are run by thousands of computational chemistry researchers around the world, on everything from ageing, dusty desktops, to the most powerful supercomputers on the planet.
It might be supposed that this would lead to a deluge of valuable data, but the surprising fact remains that most of this data, if it is archived at all, usually lies hidden away on hard disks or buried on tape backups; often lost to the original researcher and never seen by the wider chemistry community at all.
However, it is widely accepted that if the results of all these calculations were publicly accessible it would be extremely valuable as it would:

- avoid the costly duplication of results,
- allow different codes to be easily validated and benchmarked,
- provide the data required for the development of new methods,
- provide a valuable resource for data mining,
- provide an easy, automated way of generating and archiving supporting information for publications.

In the rare cases when data is made openly available, the output of calculations are inevitably produced in a code-specific format; there being no currently accepted output standard. This means that interpreting or reusing the data requires knowledge of the code, or the use of specific software that understands the output. A standard output format would:

- allow tools, (e.g. GUI's) to operate on the input and output of any code supporting the format, vastly increasing their utility and range,
- enable different codes to interoperate to create complex workflows,
- additionally, if a semantic model underlies the format, data can easily be validated.

The benefits of a common data standard and results databases are obvious, but several previous efforts have failed to address them, largely because of an inability to settle on a data standard or provide any useful tools that would make it worthwhile for code developers to expend the time to make their codes compatible.

The Quixote project aims to tackle both of these problems in a pragmatic way, building an infrastructure that can be used to both archive and search calculations on a local hard-drive, or expose the data on publicly accessible servers to make it available to the wider community.

The vision with which we started the Quixote project some months ago is one in which all data generated in computational QC research projects is used with maximal efficiency, is almost immediately made available online and aggregated into global search indexes; a vision in which no work is duplicated by researchers and everyone can get an overall picture of what has been calculated for a given system, for a given scientific question, in a matter of minutes; a vision in which all players collaborate to achieve maximum interoperability between the different stages of the scientific process of discovery, in which commonly agreed, semantically rich formats are used, and all publications expose the data as readable and reusable supplementary material, thus enforcing reproducibility of the results; a vision in which good practices are wide spread in the community, and the greatest benefit is earned from the effort invested by everyone working in the field.

With the prototype presented in this article, which has been validated by real use cases, we believe this vision is beginning to be accomplished.

Also the methodological approach in Quixote is special: The data standard will be consolidated around the tools and encourage its adoption by providing code and tool developers with an obvious reason for adopting the data standard; the "If you build it, they will come" approach. The project is rooted in the belief that scientific codes and data should be "open", and we are therefore focussing our efforts on using existing open-source solutions and standards where possible, and then developing any additional tools within the project. The Quixote project is itself completely open, de-centralised and community-driven. It

is composed of passionate researchers from around the globe that are happy to collaborate with anyone who shares our aims.

Methods

Work methodology

(I would add here all the ways in which the Quixote workflow is special: open, non-hierachic, etc.)

Quixote components

(Make here a nice summary of the whole Quixote infrastructure, maybe with a couple of nice diagrams; then we discuss each component individually)

JUMBO Converters
CML, CMLcomp and dictionaries
Lensfield
RESTful uploading and downloading
Chemp# repository
D-Space metadata repository

Authors contributions

P. Echenique has written the manuscript, participated in the design of the Quixote system and help develop some of the tools contained in it.

J. Estrada has written the manuscript, participated in the design of the Quixote system and help develop some of the tools contained in it.

(add the rest of the authors here)

Acknowledgements

We thank all the many researchers that have contributed to the work discussed here with their ideas, testing and support; particularly **(don't forget anybody here)**. We also thank the Zaragoza Center for Advanced Modeling (ZCAM), and specially its Director, Michel Mareschal, for hosting and co-organizing the vibrant workshop in which the Quixote project was born.

P. Echenique aknowledges support from the research grants E24/3 (DGA, Spain), FIS2009-13364-C02-01 (MICINN, Spain). P. Echenique and J. Estrada aknowledge support from the research grant 200980I064 (CSIC, Spain), and and ARAID and Ibercaja grant for young researchers (Spain). The mentioned meeting

has been funded by ZCAM, the University of Zaragoza, Piregrid, the Aragón Government, and the Spanish Ministry of Science and Innovation. **(please add your grants and funders here)**

References

1. Frisch MJ, et al.: **Gaussian 03, Revision C.02**. [Gaussian, Inc., Wallingford, CT, 2004].
2. Gordon MW M S ans Schmidt: **Advances in electronic structure theory: GAMESS a decade later**. In *Theory and Applications of Computational Chemistry: The first forty years*. Edited by Dykstra CE, Frenking G, Kim KS, Scuseria, Amsterdam: Elsevier 2005:1167–1189.
3. Valiev M, Bylaska EJ, Govind N, Kowalski K, Straatsma TP, van Dam HJJ, Wang D, Nieplocha J, Apra E, Windus TL, de Jong WA: **NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations**. *Comput. Phys. Commun.* 2010, **181**:1477.
4. Karlström G, Lindh R, Malmqvist PA, Roos BO, Ryde U, Veryazov V, Widmark PO, Cossi M, Schimmelpfennig B, Neogrady P, Seijo L: **MOLCAS: A program package for computational chemistry**. *Comp. Mat. Sci.* 2003, **28**:222.
5. Echenique P, Alonso JL: **A mathematical and computational review of Hartree-Fock SCF methods in Quantum Chemistry**. *Mol. Phys.* 2007, **105**:3057–3098.
6. Shao Y, et al.: **Advances in methods and algorithms in a modern quantum chemistry program package**. *Phys. Chem. Chem. Phys.* 2006, **8**:3172–3191.
7. e-SciDR: **Towards a European e-Infrastructure for e-Science digital repositories - Final Report**. *http://e-scidr.eu/* 2008.
8. ESF: **European Computational Science Forum: The "Lince Initiative": from computers to scientific excellence** 2009.
9. Harding ME, Metzroth T, Gauss J, Auer AA: **Parallel calculation of CCSD and CCSD(T) analytic first and second derivatives**. *J. Chem. Theory Comput.* 2008, **4**:64–74.
10. Jensen F: *Introduction to Computational Chemistry*. Chichester: John Wiley & Sons 1998.
11. Szabo A, Ostlund NS: *Modern Quantum Chemistry: Introduced to Advanced Electronic Structure Theory*. New York: Dover Publications 1996.
12. Pople JA: **Nobel lecture: Quantum chemical models**. *Rev. Mod. Phys.* 1999, **71**:1267–1274.
13. Echenique P, Alonso JL: **Efficient model chemistries for peptides. I. General framework and a study of the heterolevel approximation in RHF and MP2 with Pople split-valence basis sets**. *J. Comput. Chem.* 2008, **29**:1408–1422.
14. Perczel A, Jákli I, Csizmadia IG: **Intrinsically stable secondary structure elements of proteins: A comprehensive study of folding units of proteins by computation and by analysis of data determined by X-ray crystallography**. *Chem. Eur. J.* 2003, **9**:5332–5342.
15. Perczel A, Hudáky P, Füzéry AK, Csizmadia IG: **Stability issues of covalently and noncovalently bonded peptide subunits**. *J. Comput. Chem.* 2004, **25**:1084–1100.
16. Cancès E, DeFranceschi M, Kutzelnigg W, Le Bris C, Maday Y: **Computational quantum chemistry: A primer**. In *Handbook of numerical analysis. Volume X: Special volume: Computational chemistry*. Edited by Ciarlet P, Le Bris C, Elsevier 2003:3–270.
17. Feller D: **The role of databases in support of Computational Chemistry**. *J. Comput. Chem.* 1996, **13**:1571–1586.
18. Li L, Zhang R, Chen J, Zhang Y, Li L, Zhao Z: **ChemDataBase 2: An enhanced chemical database management system for virtual screening**. In *The Fifth Annual ChinaGrid Conference* 2010.
19. Řezáč J, Jurečka P, Riley KE, Černý J, Valdes H, Pluháčková K, Berka K, Řezác T, Pitoňák M, Vondrášek J, Hobza P: **Quantum chemical benchmark energy and geometry database for molecular clusters and complex molecular systems (www.begdb.com): A users manual and examples**. *Collect. Czech. Chem. Commun.* 2008, **73**:1261–1270.

20. Giese TJ, Gregersen BA, Liu Y, Nam K, Mayaan E, Moser A, Range K, Nieto Faza O, Silva Lopez C, Rodriguez de Lera A, Schaftenaar G, Lopez X, Lee TS, Karypis G, York DM: **QCRNA 1.0: A database of quantum calculations for RNA catalysis**. *J. Mol. Graph. Model.* 2006, **25**:423–433.

21. Kotochigova S, Levine ZH, Shirley EL, Stiles MD, Clark CW: **Local-density-functional calculations of the energy of atoms**. *Phys. Rev. A* 1997, **55**:191–199.

22. Murray-Rust P, Rzepa HS: **Chemical markup, XML, and the Worldwide Web. 1. Basic principles**. *J. Chem. Inf. Comput. Sci.* 1999, **39**:928–942.

23. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa HS, Steinbeck C, Wegner J, Willighagen EL: **The Blue Obelisk – Interoperability in Chemical Informatics**. *J. Chem. Inf. Model.* 2006, **46**:991–998.

24. Schuchardt KL, Didier BT, Elsethagen T, Sun L, Gurumoorthi V, Chase J, Li J, Windus TL: **Basis Set Exchange: A community database for computational sciences**. *J. Chem. Inf. Model.* 2007, **47**:1045–1052.

25. O'Boyle NM, Tenderholt AL, Langner KM: **cclib: a library for package-independent computational chemistry algorithms**. *J. Comput. Chem.* 2008, **29**:839–845.

26. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen EL: **The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics**. *J. Chem. Inf. Comput. Sci.* 2003, **43**:493–500.

Figures

Figure 1 - Sample figure title

A short description of the figure content should go here.

Figure 2 - Sample figure title

Figure legend text.

Tables

Table 1 - Sample table title

Here is an example of a *small* table in LaTeX using `\tabular{...}`. This is where the description of the table should go.

My Table		
A1	B2	C3
A2	...	..
A3	..	.

Table 2 - Sample table title

Large tables are attached as separate files but should still be described here.

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.