JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform

M. Álvarez-Moreno,*[,†,‡] C. de Graaf,[‡,∥] N. López,[†] F. Maseras,[†,§] J. M. Poblet,[‡] and C. Bo*[,†,‡]

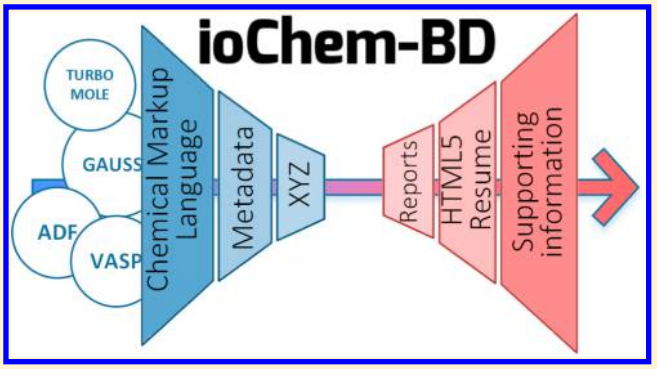
[†]Institute of Chemical Research of Catalonia, ICIQ, Av. Països Catalans 16, 43007 Tarragona, Catalonia, Spain

[‡]Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, C/Marcel·lí Domingo s/n, 43007 Tarragona, Catalonia, Spain

[§]Department of Chemistry, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain

[∥]Catalan Institution for Research and Advanced Studies, ICREA, Passeig Lluis Companys 23, 08010 Barcelona, Catalonia, Spain

ABSTRACT: We present the **ioChem-BD** platform (www.iochem-bd.org) as a multiheaded tool aimed to manage large volumes of quantum chemistry results from a diverse group of already common simulation packages. The platform has an extensible structure. The key modules managing the main tasks are to (i) upload of output files from common computational chemistry packages, (ii) extract meaningful data from the results, and (iii) generate output summaries in user-friendly formats. A heavy use of the Chemical Mark-up Language (CML) is made in the intermediate files used by **ioChem-BD**. From them and using XSL techniques, we manipulate and transform such chemical data sets to fulfill researchers' needs in the form of HTML5 reports, supporting information, and other research media.



1. INTRODUCTION

Intensive high performance computing is one of the pillars to accelerate materials discovery and development in many fields of science and engineering, most prominently chemistry, physics, and related areas. The volume of information generated daily, coming from the results of scientific calculations, is increasing exponentially. For instance, in our lab, scientists (a group of about 10) generate 1.3TB weekly. Its conservation on physical media is favored right now by the cheap price of storage per bit of information as well as an increase in the available telecommunications infrastructure bandwidth. This fact makes the public and private centers generate and store more and more terabytes of information. In our particular case, this information corresponds to the outcome of calculations. At present, storage of computational simulations has not been identified as a bottleneck in most data and supercomputing centers. However, the main global players in the Internet business have already introduced the "Big Data"[1] concept to start looking for solutions to maintain all physical data storage systems sustainable and provide convenient access. Solutions based on what is called "the cloud" along with concepts derived from the "social networks" are leading to a reformulation of how and in what physical space the information shall be stored. As a result, there is a growing demand for tools to order the storage, allow analysis, and simplify the presentation of significantly large volumes of growing data in an amenable, transparent, and reliable manner.[2]

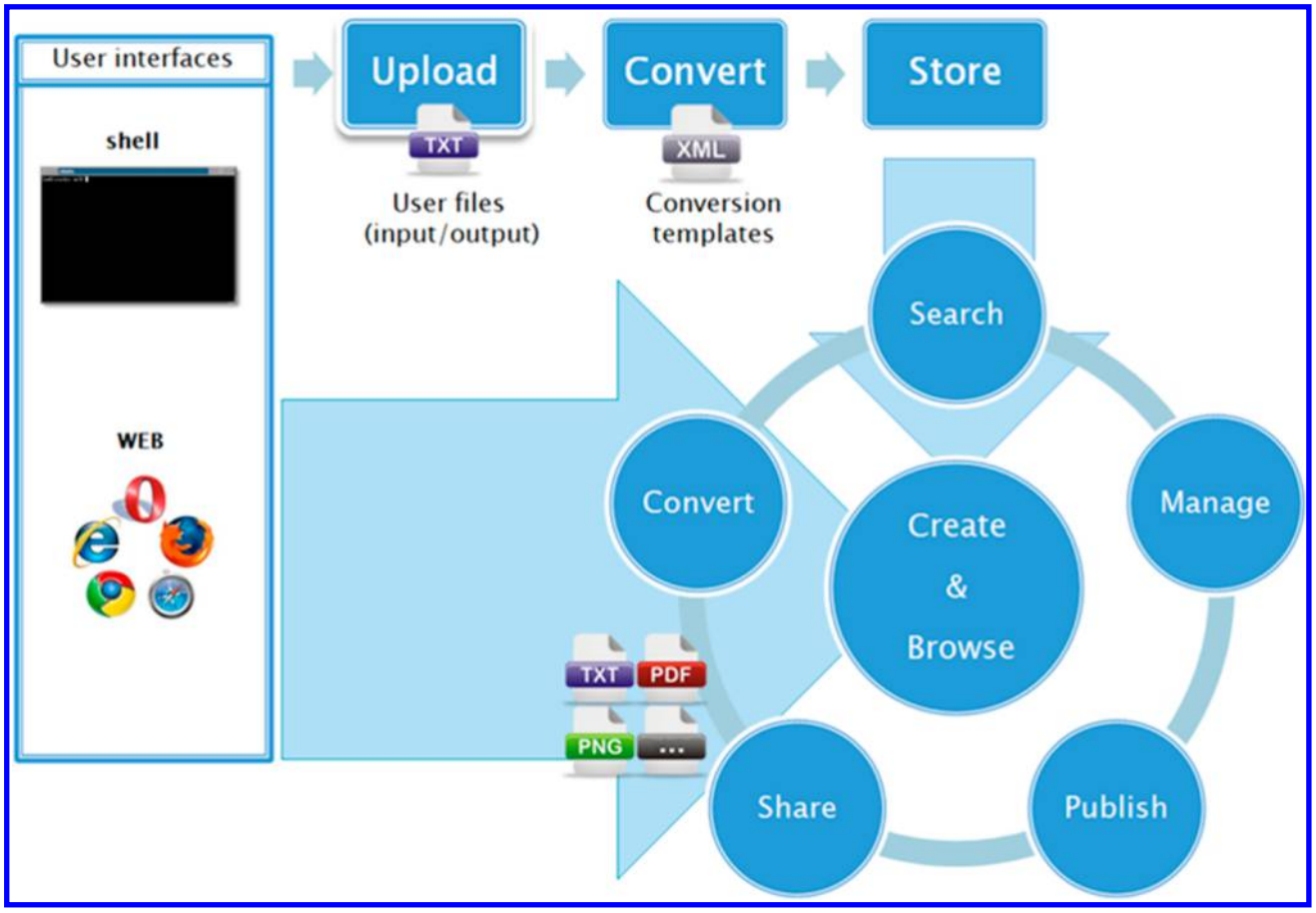
Only a very small percentage of the information currently stored in data centers is hierarchically indexed.[3] This simply means that the information available is impossible to process; the information bits are hardly usable by any person other than the creator himself. Tim Berners-Lee, one of the original creators of the World Wide Web, identified the need to transform numerical data into "raw data".[4] This raw data is the desirable state where information is meaningful because it is enriched with labels that contextualizes it, the labels being "metadata". Once contextualized, searching the ocean of information is more efficient. Moreover, the search process becomes a process of knowledge creation, as it is possible to establish new connections, in what is already called "linked data".[5]

The application of these concepts to the field of computational chemistry is hindered by the heterogeneity of the packages used in the atomistic simulations of molecules and materials. As a result, the outcome of atomistic simulations addressing chemical or physical problems and based on the application of the Schrödinger equation are presented in a disperse manner, with sparse data showing only some of the key aspects as geometries, energies, and chemical and/or physical properties. The wave function or density data does not belong to this set due to the following reasons: (i) the size of these files is excessive for our purposes; (ii) the speed of the

Received: September 30, 2014
Published: December 3, 2014

ACS Publications © 2014 American Chemical Society

Scheme 1. ioChem-BD System Overview[a]



[a]Web nature of the CREATE and BROWSE modules is highlighted. Both modules share functionalities like searching and browsing chemical datasets, exporting to third party formats, and publishing results, but without losing sight of the original private−public sense of each module.

computers, once the optimized geometries are available, allows the recreation of this massive data. The ultimate consequence is that the data published in the scientific journals of chemistry, physics, nanoscience, biochemistry and related areas are not homogeneous; they are often incomplete, are hardly consulted in bulk, and are rarely reused. The high degree of diversity in the data formats requires the definition of standards.[6] There have been technological initiatives[7] like the Quixote project that implement solutions on multiple aspects of this problem like data format unity and data management.[8] Parallel initiatives to ours are under development,[9] for instance, the AiidA software, that according to the available documentation focuses on elaborated workflows to carry out complex calculations in an automated manner. Other purpose-dedicated databases have been generated by the groups of MIT, Berkeley,[10] and Standford.[11]

In this paper, we present an alternative platform, **ioChem-BD**, encompassing a variety of aspects in the definition of standards for treatment, hierarchical storage, and retrieval of data. Our platform automates the extraction of relevant data and its conversion into fully tagged information in a distributed database. It provides tools for the researcher to validate, enrich, publish, and share information, and tools in the cloud to access it and view it.

2. BASE TECHNOLOGIES

The keystone in the definition of the project is that it employs high reliability software technologies widely used in the Internet world that are extended to cover our particular area of interest. We chose eXtensible Markup Language (XML)[12] as the container element of all information for its reliability, format neutrality, and ease of validation (using XSD verification tools).[13] To be more specific, we chose Chemical Markup Language (CML) implementation because it contains all semantics necessary to describe most chemical entities.[14] With calculations in CML format, querying its content for specific information is extremely easy and efficient by using XPATH queries.[15] Working with XML provides a wide range of conversion operations from CML files into any other existing or future format using eXtensible Stylesheet Language Transformations (XSLT).[16]

As access to information is becoming more universal, the data in the system is reachable through the Internet by any digital device on the market, following the latest existing Web standards in communication.[12,17] Users have at their disposal the latest search,[18,19] display,[20] and data-labeling tools,[14] and also, there exists communication channels enabled to propose new features.

In terms of data storage, information is distributed among the content generators, creating a mesh topology in which the service is always available and accessible on the network. Being a cloud system, it has the necessary standards in data

dx.doi.org/10.1021/ci500593j | J. Chem. Inf. Model. 2015, 55, 95−103

definition[21−23] in order to connect to other digital repositories and external Web services to build up a network of interconnected semantic data to provide the most sense to the user experience.

Finally, to enhance industrial implementation, the platform allows secure[24,25] and reliable channels for the communication and collaboration between users, groups, and/or centers but with the highest privacy standards for third partners. All information has configurable levels of access and licensing, allowing for adaptation to the specific legal needs of each entity.

Architecture. The **ioChem-BD** platform is composed of two main modules that work independently, labeled CREATE and BROWSE. Both of them are executed as Java Web services.

The CREATE module is designed to extract the information from the output files generated by the computational chemistry packages and store it in an organized way. It currently manages output files from the programs Gaussian,[26] ADF,[27] and VASP.[28] The list of accessible codes should be expanded in the midterm future to codes such as SIESTA,[29] TURBOMOLE,[30] MOLCAS,[31] and ORCA.[32] The CREATE module shall be used by the scientists that execute the simulations (creator).

The BROWSE module is designed as a tool to explore and use the data contained in the database, and it resides in the cloud. The BROWSE module has a much broader scope and is useful to researchers interested in accessing the computational Big Data.

The combination of both modules gives the scientific community storage and access to all the Chemistry Big Data in a way schematically shown in Scheme 1.

CREATE Module. The CREATE module contains two different subunits that allow: (i) the extraction and structuring of the relevant output data and (ii) the publication of such data and derivates into BROWSE.

CREATE Module: Data Extraction from Output Files. The system initially works with input and output files from the computational chemistry packages described above. In Gaussian and ADF, the results are stored in a single file that needs to be extracted. However, for other computational codes, the CREATE module needs a group of files. This is the particular case for VASP calculations, where relevant data are split into multiple files from inputs like POTCAR or output files OUTCAR (summary), CONTCAR (geometries), XDATCAR (trajectory), and vasprun.xml. The **ioChem-BD** platform outputs a single CML file from these sources. The upload of these files to the CREATE module is done inside a layered process where we simultaneously parse and tag all relevant data; we infer its metadata and capture the molecular geometry. A scriptable shell upload utility is used to do file conversion and upload calculations straight from HPC clusters to the CREATE module; there is also an alternative mechanism to upload content from the user Web browser. A detection algorithm decides which format extraction templates to use based on the calculation file content. Once the file format is elucidated, the appropriate templates for such formats are selected, and the first conversion is performed from plain text to CML using a modified version of the JUMBOConverters library.[33] This is followed by a second conversion to reorder CML tags so they comply with the CompChem convention.[34] This process can be used on individual or multiple output files as is depicted in Figures 1 and 2.

As an additional feature, the module is able to attach directly to other supporting files such as calculation input files, graphics, and text, and all needed associated gray literature. These

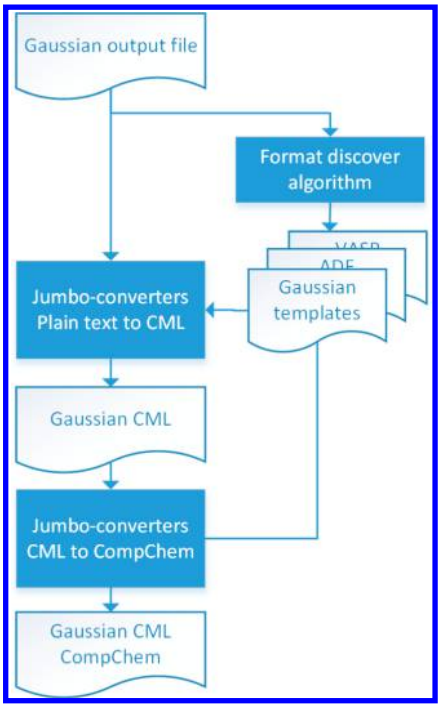


Figure 1. Conversion workflow for individual output files. This two-step process is the default behavior for file conversion inside the JUMBOConverters library, from output files into CML elements, and then to compliant CML CompChem.

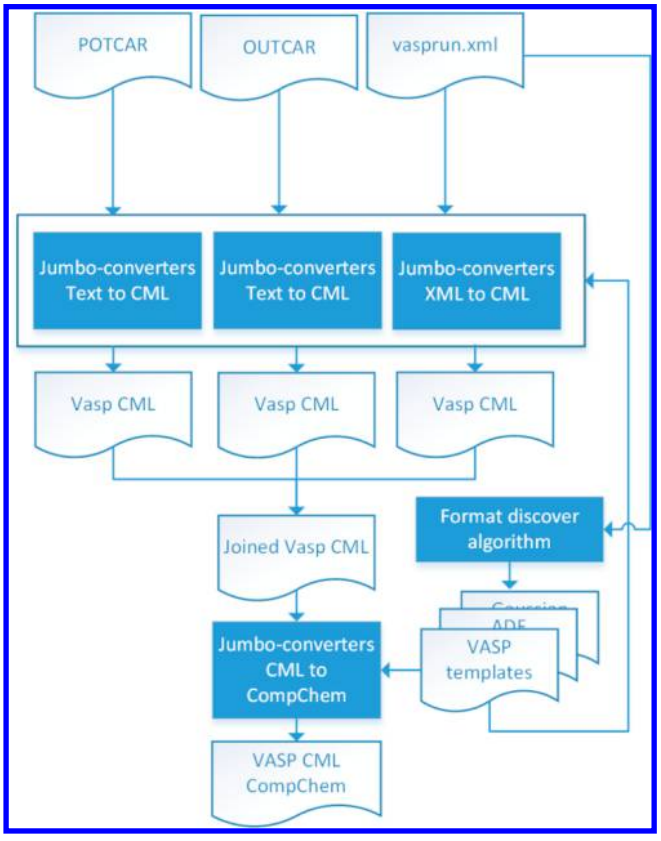


Figure 2. Conversion workflow for multiple output files. Our customized JUMBOConverters library accepts multiple output files for its unification into a single CML file. Uploaded files can be a mixture of plain text and XML files.

dx.doi.org/10.1021/ci500593j | J. Chem. Inf. Model. 2015, 55, 95−103

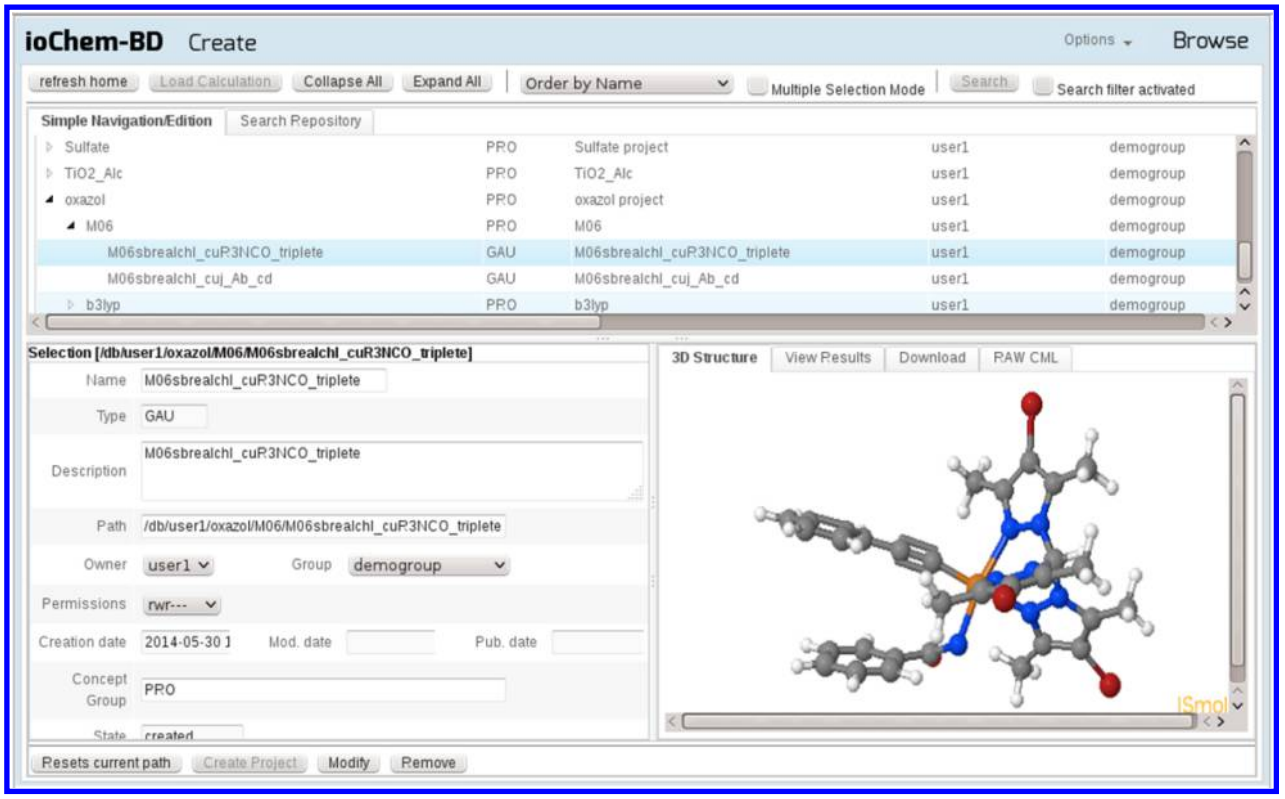


Figure 3. CREATE main panel view. The hierarchical tree in the upper section allows browsing all uploaded content. Selecting an element from it will fill lower panels with more detailed information and its available display actions.

additional files are not processed, so the future user of the database should process them him/herself. Such files will be paired with the calculation CML file during its existence inside the **ioChem-BD** system to provide further information on the calculation (Figure 3).

Once the CompChem CML file is generated, a second data flow is triggered to extract the corresponding metadata fields. By using XSLT style sheets, we infer fields such as type of calculation, methods used, basis set, charge, multiplicity, and several others. From the CREATE database, we will also retrieve additional information such as structural (which files are involved in this upload process) and administrative (how these files were generated). Figure 4 depicts this process that ends building a METS compliant file of administrative, descriptive, and structural metadata containing all aspects of the upload.

Prior to the data storage by the CREATE module, there is a final step aimed to extract the final geometry, a key point to repeat the calculation if needed. This particular point sets our computational database close to other structural databases like the CSD (molecules)[35] and for the structures of compounds in crystallography COD.[36] In the case of geometry optimizations, a large number of geometries can appear in the same file. Again, we rely on XSLT templates to contain the necessary logic to retrieve the optimized geometry. The geometry is then indexed with ChemAxon JChemBase software for future substructure searches.[37]

When all these processes are completed, newly uploaded calculations are accessible on the CREATE module via tree browsing or search (Figure 3). At this point, users can browse their uploaded content. Selecting a calculation opens an auxiliary window in lower right corner with all available actions: visualize molecule on JSmol viewer,[38] view an HTML5

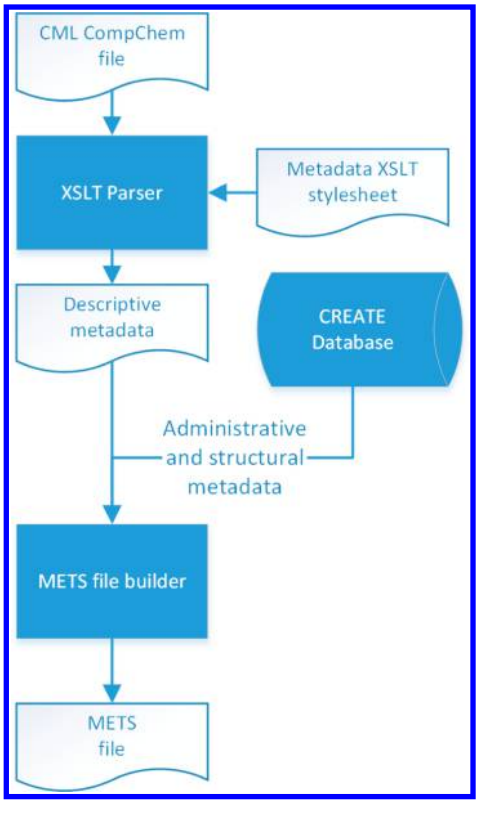


Figure 4. METS file generation workflow. Using XSLT stylesheets, we extract calculation descriptive metadata fields. Together with these fields, we append structural and administrative information to compose a METS file that fully describes our new uploaded result.

dx.doi.org/10.1021/ci500593j | J. Chem. Inf. Model. 2015, 55, 95−103

resume of relevant calculation data (or other attaches files), download and visualize CML, and attached files. To keep the system's extensibility, all actions applicable to content are implementations of an abstract *Action* class that is managed via an *ActionManager* object. Such a class acts also as a class loader. This allows upgrading the system with new calculation operations without the need to update its code, just dropping a new *Action* implementation class package in the Web server class path.

Once uploaded, it is possible to search the stored data. The search functionality relies on standard database queries in conjunction with the JChemBase search engine to filter its content.[37] As shown in Figures 5 and 6, users can query

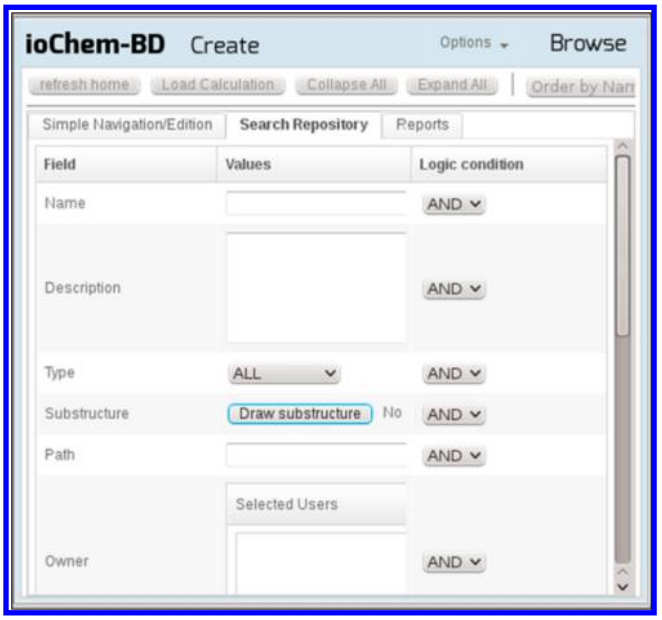


Figure 5. CREATE search panel allows users to define multiple search criteria using boolean logic. Such queries range from administrative metadata, chemical related terms, and chemical substructures.

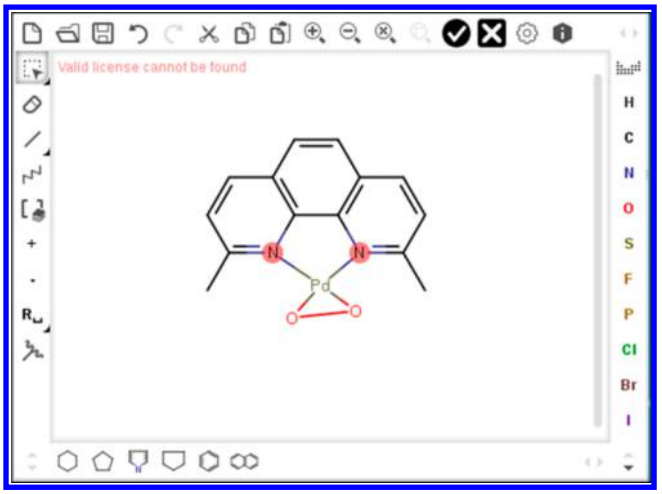


Figure 6. Search output can be narrowed by the definition of a molecular substructure that will refine its results. A visual HTML5 molecular editor is displayed on the user's browser to sketch fragments or the entire molecule.

administrative and descriptive metadata fields and use a molecular editor to sketch substructures that will be used as a search filter. Results vary depending on the privileges that the user possesses toward CREATE calculations. They are defined by fine-grained access rules set at the user, group, and others levels, like UNIX system file rights.

Next to JSmol visualization, another remarkable action is the HTML resume (Figure 7). Using XSLT style sheets, **ioChem-BD** is able to generate a fully compliant HTML5 resume that implements features such as one page presentation, all data sets exportable to other formats, compact drop-down content, device responsiveness, and its most valuable feature of being fully customizable with new data fields without the need to upgrade the platform.

Figure 7. Every uploaded calculation has a group of actions associated with it. One of them is a HTML summary that displays its most remarkable fields. Such a summary can be customized to fulfill researchers' needs and to adapt to future requirements.

Another feature delivered in HTML5 reports is the visual representation of data. A reference to Highcharts (a Javascript charting library)[39] has been included in all generated reports. This inclusion eases the process to convert plain data into interactive visual elements using (among others) line, scatter, or column charts. This inclusion behavior is easily replicable to the innumerable third party plugins that exist today in the chemistry field.

In addition, this report file can contain other rich content objects such as third party plugins, navigable data tables, and interactive graphics, among others. An example of **ioChem-BD**'s pluggability with external tools can be observed in the integration process done inside the HTML5 report generation

dx.doi.org/10.1021/ci500593j | J. Chem. Inf. Model. 2015, 55, 95−103

engine with JCAMP-MOL IR Spectrum Viewer applet.[40] During the development of this engine, there was a need to include an IR viewer so that calculated vibrational frequencies could be displayed as an additional visual field inside the resume. To do so, a java servlet was created to convert CML calculations to Jcamp-DX[41] compliant output text by the use of XSLT transformations. Now, calling this servlet with a calculation ID will return its vibrational information in the Jcamp-DX format, so appending the applet tag calling this servlet inside our report did the job. No major code development was required.

CREATE Publication Mechanism. Communication between both **ioChem-BD** modules is currently unidirectional, from CREATE to BROWSE modules, through a process called "Content publication". Publishing allows importing single calculations or groups of them to the BROWSE module to generate assets like reports. To complete this step, it is only necessary to name calculations and mark them for publication. The remaining process, REST API communications, is invisible to the user. Because both modules are written as Java Web services, the publication mechanism is done via servlets ,and published files are bundled in DSpace METS SIP[42] format during its ingestion in the BROWSE module.

From this step onward, published calculations will be called "items". As a result of the publication process, a group of URL handles referring to published items are presented. These links point to public HTML pages in the BROWSE module with the following content: (i) final calculation geometry visualization with Jsmol, (ii) expandable summary of the item's metadata, (iii) summary of the most relevant data in HTML5 format, (iv) list of downloadable content such as input files, and (v) support files and gray literature associated with calculations. In this, final content administrative metadata related to the purpose of the calculation, methodology, or other relevant information from the user acting as creator can be uploaded. Most of these sections can be mapped to CREATE Actions as they share the same conversion style sheets. Therefore, results share coherency in both modules.

BROWSE Module. The BROWSE module consists of a heavily modified version of the DSpace digital repository.[18] It has been adapted to fulfill our requirements, mainly in quantum chemistry data representation and in external services communication. Some workflows have been copied from the CREATE module to have a similar behavior between them. One of the main features in BROWSE (DSpace) instances is that they can communicate between them using the OAI-PMH protocol[21] to share item metadata. This allows building a public distributed network of theoretical chemistry and materials science repositories, which will be a great advance in terms of information socialization.

The module works by default with the Dublin Core metadata schema,[43] which is good to capture the most basic bibliographic information about any digital asset but cannot hold the description of quantum chemistry documents. However, this module is versatile enough to expand its metadata schemas with new ones, so we have created a schema focused on the computational chemistry field. Among other interesting features, the BROWSE module accepts browsing and searching content,and such content can be embargoed, exported, or syndicated depending on the users' needs.

A notable aspect of the BROWSE module is its ability to display supporting information and other derived chemical reports built with the CREATE module. As a brief overview, supporting information documents are normally composed of one or several chemical structures (normally with the XYZ format) from a series of related calculations. It can also contain extra information fields such as final energies, vibrational frequencies, spin angular momentum, etc. Supporting information documents are normally stored on heterogeneous locations like public ftp servers, private Web servers, cloud storage services, etc., depending on the data publication policy of each research center. These documents are later pointed to by journal papers as additional information related to research. Usually, they are generated manually in a tedious, time-wasting, and error-prone action that sometimes derivates on unportable digital documents (one, at most two stars of five in the Open Data Scheme).[44] We try to remove such an ineffective procedure using the supporting information generator that is integrated inside the CREATE module and whose results are displayed in BROWSE (Figure 8). It uses XSL-FO, an open format object definition language, as a bridge between raw data and multiformat output. Such a report engine is feed with CML calculations from a user selection at the CREATE main panel tree. After setting them in session, the user chooses to create a new report from it. In this case, we choose Supporting Information as report type. A fast XPath query will return

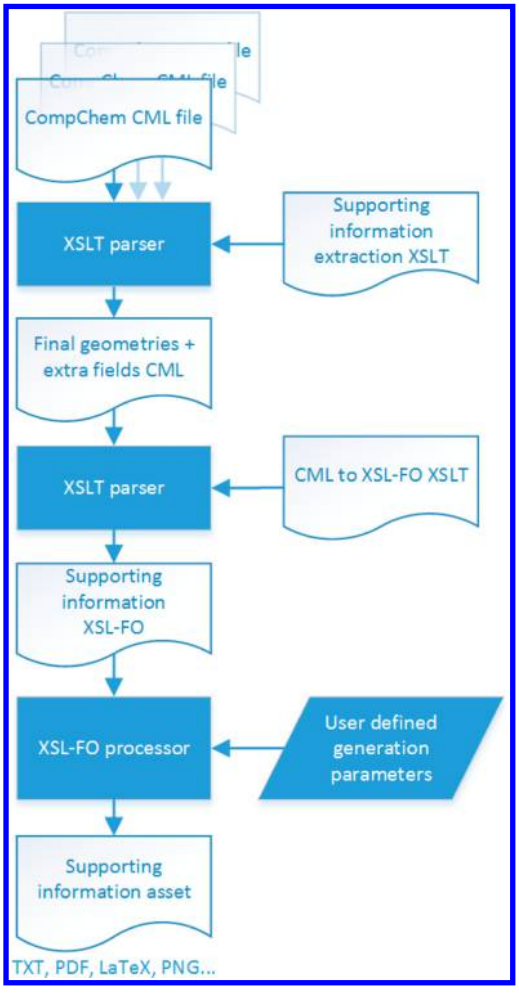


Figure 8. Supporting information report generation workflow. Starting from a user selection of calculations, the module extracts its molecular geometry (among other fields like final energies) to bundle them into a single XML file. Following iterations will convert it to a XSL-FO format and then to the user's desired output format.

dx.doi.org/10.1021/ci500593j | J. Chem. Inf. Model. 2015, 55, 95−103

additional fields (like final energies) that exist among these calculations and that will (dis)activate additional report generation options.

After setting up the report fields, the engine will extract the final geometries and other fields from chosen calculations. Then, they are joined into a single CML file. The next step in report generation is to convert CML to a XSL-FO document; with some more XSLT work, we obtain a XSL-FO document ready to be converted based on users' choice to any kind of digital document such as PDF, TXT, CSV, etc.

Using a similar process, **ioChem-BD** is able to build a daily growing set of reports. In this case, we can opt to generate two types of outputs: a ready to download multiformat XSL-FO document (similar to a supporting information report) or an HTML5 Web page that will pop up in a new tab. This last option is extremely versatile because it opens the door for adding third party plugins and other dynamic content to our report, a more powerful way to display results.

As an example of this functionality, we describe energy reaction profile report generation. CREATE users need to select a group of calculations and define a set of formulas that constitute the energy steps. The report engine will build a dynamic device-responsive HTML5 report in our browser displaying an energy profile chart for such calculations (Figure 9).

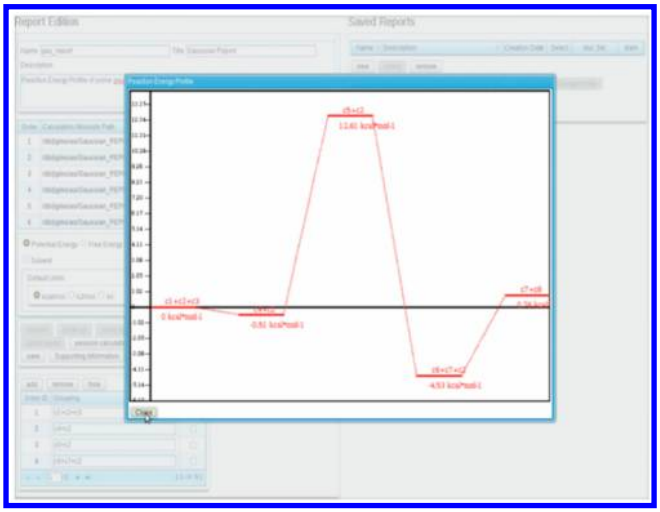


Figure 9. Example of a dynamically generated report. On the basis of user calculation selection and the definition of multiple energy reaction formulas, our platform is able to build and output reaction energy profile reports.

In terms of programming code, there is an abstract class defined for *Reports*. So new classes can implement its functions, and the *ReportManager* class will load them, appending new report types dynamically with no need to alter our existing code. Extensions to more complex outputs like R language code snippets[45] or Jmol scripts[46] are envisaged.

Inside **ioChem-BD**, all content derived from calculations is built under demand and then streamed to the user's browser. There is minimal performance loss using this dynamic generation approach, but we enormously reduce disk space requirements and increase data veracity, avoiding the massive storage of formatted content that over time can become outdated or partial.

Current developments in **ioChem-BD** are focused on the publication in the BROWSE module of calculation reports. At present, reports can only be generated in the CREATE module, but in the near future, it will be possible to generate a public handle inside BROWSE that points to a report generation page that, depending on its URL parameters, outputs its results in multiple formats.

System Adaptability and Safety Considerations. Dynamic data definition and capture is a requirement of today's chemistry computational sector. Quantum software vendors periodically release new versions of its products with the addition of new functionalities, bug fixes, data representation changes, new chemical properties, calculation methods, or atom basis, and on the other side, chemist software users demand more analysis tools and higher levels of data representation.

This constant flow of structural and representational data changes defines a list of requirements that our software tries to fulfill with loosely coupled data management rules. With our customized JUMBOConverters library, we can expand our data capture rules just by expanding the XML templates definition. We can also modify metadata capture and data presentation to the final user with the modification of inner XSLT style sheets. Mastering the skills necessary to modify and expand these rules presents a small learning curve because they are based on open and well-documented standards. Therefore, every research group can easily adapt its **ioChem-BD** instance to its requirements without the need of an external programmer.

A user authentication mechanism has been implemented with the Jasig CAS SSO Server.[25] Its session management service allows us to append new independent Web services in a modular fashion without the need to implement user credential management inside our modules.

In **ioChem-BD**, data processing documentation has the same relevance as the processes it tries to describe. An outdated documentation on a highly dynamic system as the **ioChem-BD** environment will unavoidably lead to confusion. Users cannot track down recent changes, and the reimplementation of already existing extraction rules becomes hard to avoid. In addition to this, such rules are defined on XML, a cryptic language that does not help its reading unless it is converted to a user-friendly format. These new requirements led us to develop a toolkit that manipulates Jumbo capture templates to build a SGML/XML DocBook fileset.[47] We use it as a neutral format bridge for its later conversion into a hierarchical group of Web pages in WebHelp format. The documentation generation process is triggered on every template modification and becomes instantly accessible to all CREATE users for its reference. This effectively avoids that the documentation becomes outdated.

All content managed inside **ioChem-BD** is under access control, even published items. In the CREATE module, calculation content is restricted at the user/group/others levels. In the BROWSE module, content can define fine-grained access rules and also set content embargos depending on third party publication requirements. Splitting the system into two separated modules that should be installed on separated Web servers increases the overall security of the system. The CREATE module will hold internal research data and should be deployed in internal Web servers with few open ports to capture uploaded calculations from HPC and for the publication mechanism. The BROWSE module can be moved to a public Web server, where published items will reside and

also will be referred to by their handles. The whole system relies on HTTPS protocol for its communication among users and modules to ensure that data is always encrypted when transferred. There is an "additional" CAS module in charge of user validation that uses tokens for single sign on/single sign off session management, which greatly simplifies the session management code and detaches it from our modules.

3. CONCLUSIONS

The massive use of simulation techniques in chemical research generates huge amounts of information, known as "the Big Data problem". The main obstacle for managing enormous volumes of information concerns its storage in such a way that facilitates data mining as a strategy to optimize the processes that allow scientists to face the challenges of sustainability, knowledge, and the rational use of existent resources. We created **ioChem-BD** (www.iochem-bd.org) as a group of services in the cloud to manage computational chemistry input and output files. As with other database-related projects, the concepts underlying our platform rely on well-defined standards, and it manages treatment, hierarchical storage, and data recovery tools to facilitate data mining. This software implements new methodological strategies that promote optimal reuse of results and accumulated knowledge and that improve researchers' daily productivity. It automates the extraction of relevant data and transforms numerical data into tagged data inside its database. This platform provides tools for the researcher in order to validate, enrich, publish, and share information, and tools for accessing and visualizing data. Other modules allow the automatic creation of both reaction energy profile plots (by combining data of a set of molecular entities) and Supporting Information files. Besides, **ioChem-BD** is capable of performing kinetic analysis from reaction energy profiles, QSSR analysis, or build data sets for screening, for instance. Evaluation of these facilities is currently being carried out in our groups.

The final goal is to build a new reference tool in computational chemistry research and to fill the gap between the generation of results and the publication of manuscripts embedded in bibliography management and services to third parties. Future implementations will include integration with a semantic database by taking advantage of XSLT transformations to create data triples of every uploaded calculation. With such information, we will be in the position to connect our semantic data with other external data sources and to develop a REST API to open bridges between the BROWSE module and third party data services.[44]

A list of current working instances of **ioChem-BD** software and a demo server are accessible at www.iochem-bd.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: moises.alvarez@urv.cat (M.Á-M.).
*E-mail: cbo@iciq.cat (C.B.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Financial support for this work from the AGAUR (ref 2009 SGR 25, 2014 SGR 199, and 2014 SGR 409) of Generalitat de Catalunya is grateful acknowledged, along with the Spanish Ministry of Science and Innovation (project CTQ2011-29054-C02-01/BQU; CTQ2011-29054-C02-02/BQU; CTQ2011-27033/BQU; CTQ2012-3382/BQU; CTQ2011-23140). We also thank MINECO for support through Severo Ochoa Excellence Accreditation 2014-2018 (SEV-2013-0319). COST Action CM1203 "Polyoxometalate Chemistry for Molecular Nanoscience (PoCheMoN)", COST Action "ECOSTBio CM1305", and ERC-2010-258406 are also gratefully acknowledged.

REFERENCES

(1) Lynch, C. Big data: How do your data grow? *Nature* **2008**, *455*, 28−29.
(2) Harvey, M. J.; Mason, N. J.; Rzepa, H. S. Digital data repositories in chemistry and their integration with journals and electronic notebooks. *J. Chem. Inf. Model.* **2014**, *54*, 2627−2635 DOI: 10.1021/ci500302p.
(3) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, DOI: 10.1038/sdata.2014.22.
(4) Berners-Lee, T. The Next Web. Ted Conference. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html (accessed September 17, 2014).
(5) Frey, J. G.; Bird, C. L. Cheminformatics and the semantic Web: Adding value with linked data and enhanced provenance. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 465−481.
(6) Phadungsukanan, W.; Kraft, M.; Townsend, J.; Murray-Rust, P. The semantics of chemical markup language (CML) for computational chemistry: CompChem. *J. Cheminf.* **2012**, *4*, 15.
(7) Chen, M.; Stott, A. C.; Li, S.; Dixon, D. A. Construction of a Robust, large-scale, collaborative database for raw data in computational chemistry: The collaborative chemistry database tool (CCDBT). *J. Mol. Graphics Modell.* **2012**, *34*, 67−75.
(8) Adams, S.; de Castro, P.; Echenique, P.; Estrada, J.; Hanwell, M. D.; Murray-Rust, P.; Sherwood, P.; Thomas, J.; Townsend, J. The Quixote Project: Collaborative and open quantum chemistry data management in the Internet age. *J. Cheminf.* **2011**, *3*, 38.
(9) AiiDA Project Home Page. http://www.aiida.net/ (accessed November 17, 2014). Computational Materials Repository (CMR) Home Page. http://cmr.fysik.dtu.dk (accessed September 17, 2014). Novel Materials Discovery Repository (NoMaD) Home Page. http://nomad-repository.eu (accessed September 17, 2014). CCSIRO Nanostructure Data Bank Home Page. https://data.csiro.au/dap/search?q=nanostructure (accessed September 17, 2014).
(10) The Materials Project Home Page. https://www.materialsproject.org (accessed September 22, 2014).
(11) Hummelshøj, J. S.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nørskov, J. K. CatApp: A Web application for surface chemistry and heterogeneous catalysis. *Angew. Chem., Int. Ed.* **2012**, *51*, 272−274.
(12) World Wide Web Consortium. Extensible Markup Language (XML) 1.0, third edition, specification. http://www.w3.org/TR/REC-xml (accessed September 17, 2014).
(13) Java schema validation class, javadoc definition. http://docs.oracle.com/javase/7/docs/api/javax/xml/validation/Validator.html (accessed September 17, 2014).
(14) Adams, N.; Cannon, E. O.; Murray-Rust, P. Chemaxiom−An ontological framework for chemistry in science. *Nat. Proc.* **2009**, DOI: 10.1038/npre.2009.3714.1.
(15) World Wide Web Consortium. XML Path Language, Version 1.0 http://www.w3.org/TR/xpath (accessed September 17, 2014).
(16) World Wide Web Consortium. XSL Transformations (XSLT), Version 1.0, W3C Recommendation, November 16, 1999. http://www.w3.org/TR/xslt (accessed September 17, 2014).
(17) HTML5 − A Vocabulary and Associated APIs for HTML and XHTML. http://www.w3.org/TR/2012/CR-html5-20121217/ (accessed September 17, 2014).
(18) Smith, M.; Barton, M.; Bass, M.; Branschofsky, M.; McClellan, G.; Stuve, D.; Walker, J. H. DSpace: An Open Source Dynamic Digital

Repository. *D-Lib Magazine* **2003**, http://www.dlib.org/dlib/january03/smith/01smith.html.
(19) Apache Lucene. A high-performance, full-featured text search engine library. http://lucene.apache.org (accessed September 17, 2014).
(20) Jmol Home Page. http://jmol.sourceforge.net/ (accessed September 17, 2014).
(21) Lagoze, C.; Van de Sompel, H. The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, New York, U.S.A., 2001.
(22) Gartner, R. *METS: Metadata Encoding and Transmission Standard*; JISC Techwatch Report TSW; Library of Congress: Washington, DC, October 2−5, 2002.
(23) Allinson, J.; François, S.; Lewis, S. Sword: Simple Web-service offering repository deposit. *Ariadne* **2008**, http://www.ariadne.ac.uk/issue54/allinson-et-al/.
(24) HTTP over TLS Description. https://tools.ietf.org/html/rfc2818/ (accessed September 17, 2014).
(25) Addison, M. S.; Battaglia, S.; Petro, A. Jasig CAS Documentation. http://jasig.github.io/cas/4.0.0/index.html (accessed September 17, 2014).
(26) Gaussian Home Page. http://jasig.github.io/cas/4.0.0/index.html (accessed September 17, 2014).
(27) ADF Home Page. http://www.scm.com/ADF (accessed September 17, 2014).
(28) VASP Home Page. http://www.vasp.at (accessed September 17, 2014).
(29) SIESTA Home Page. http://departments.icmab.es/leem/siesta (accessed September 17, 2014).
(30) Turbomole Home Page. http://www.turbomole.com (accessed September 17, 2014).
(31) Molcas Home Page. http://www.molcas.org (accessed September 17, 2014).
(32) Orca Home Page. http://cec.mpg.de/forum (accessed September 17, 2014).
(33) JUMBOconverters. Main Project Page. https://bitbucket.org/wwmm/jumbo-converters (accessed September 17, 2014).
(34) Murray-Rust, P.; Townsend, J.; Adams, S. E.; Phadungsukanan, W.; Thomas, J. The Semantics of chemical markup language (CML): Dictionaries and conventions. *J. Cheminf.* **2011**, *3*, 43.
(35) Cambridge Structural Database Home Page. http://www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/CSD.aspx (accessed September 17, 2014).
(36) Crystallography Open Database Home Page. http://www.crystallography.net/ (accessed September 17, 2014).
(37) JChem Base, Chemical Interface to Relational Database Engines. http://www.chemaxon.com/products/jchem-base (accessed September 17, 2014).
(38) JSmol, Sourceforge Project. http://sourceforge.net/projects/jsmol/ (accessed September 17, 2014).
(39) Highcharts Home Page. http://www.highcharts.com (accessed September 17, 2014).
(40) Hanson, R. M.; Lancashire, R. J. In *JCAMP-MOL: A JCAMP-DX extension to allow interactive model/spectrum exploration using Jmol and JSpecView*. The ACS 2013 Symposium on Exchangeable Data Formats,American Chemical Socoety, September 11, 2013, Indiana, IN, U.S.A.
(41) IUPAC CPEP Subcommittee on Electronic Data Standards Home Page. http://www.jcamp-dx.org (accessed September 17, 2014).
(42) DSpace METS Document Profile for Submission Information Packages (SIP). https://wiki.duraspace.org/display/DSPACE/DSpaceMETSSIPProfile (accessed September 17, 2014).
(43) DCMI Metadata Terms definition Pge. http://dublincore.org/documents/dcmi-terms/ (accessed September 22,2104).
(44) Five Star Open Data Home Page. http://5stardata.info (accessed September 17, 2014).
(45) The R Project for Statistical Computing. http://www.r-project.org (accessed September 17, 2014).
(46) Jmol /JSmol Interactive Scripting Documentation. http://chemapps.stolaf.edu/jmol/docs (accessed September 17, 2014).
(47) Ortiz, I. M.; Moreno, P.; Sierra, J. L.; Manjón, B. F. Using DocBook and XML technologies to create adaptive learning content in technical domains. *Int. J. Comput. Sci., Appl.* **2006**, *3*, 91−108.