4-2007

Basis Set Exchange: A Community Database for Computational Sciences

Karen L. Schuchardt
Pacific Northwest National Laboratory

Brett T. Didier
Pacific Northwest National Laboratory

Todd Elsethagen
Pacific Northwest National Laboratory

Lisong Sun
Pacific Northwest National Laboratory

Vidhya Gurumoorthi
Pacific Northwest National Laboratory

See next page for additional authors

Basis Set Exchange: A Community Database for Computational Sciences

Abstract

Basis sets are some of the most important input data for computational models in the chemistry, materials, biology, and other science domains that utilize computational quantum mechanics methods. Providing a shared, Web-accessible environment where researchers can not only download basis sets in their required format but browse the data, contribute new basis sets, and ultimately curate and manage the data as a community will facilitate growth of this resource and encourage sharing both data and knowledge. We describe the Basis Set Exchange (BSE), a Web portal that provides advanced browsing and download capabilities, facilities for contributing basis set data, and an environment that incorporates tools to foster development and interaction of communities. The BSE leverages and enables continued development of the basis set library originally assembled at the Environmental Molecular Sciences Laboratory.

Disciplines

Chemistry

Comments

Reprinted (adapted) with permission from *Journal of Chemical Information and Modeling* 47 (2007): 1045, doi:10.1021/ci600510j.

Authors

Karen L. Schuchardt, Brett T. Didier, Todd Elsethagen, Lisong Sun, Vidhya Gurumoorthi, Jared Chase, Jun Li, and Theresa Lynn Windus

Basis Set Exchange: A Community Database for Computational Sciences

Karen L. Schuchardt,*,† Brett T. Didier,† Todd Elsethagen,† Lisong Sun,† Vidhya Gurumoorthi,† Jared Chase,† Jun Li,† and Theresa L. Windus‡

Pacific Northwest National Laboratory, Richland, Washington 99352, and Department of Chemistry, Iowa State University, Ames, Iowa 50010

Received November 13, 2006

Basis sets are some of the most important input data for computational models in the chemistry, materials, biology, and other science domains that utilize computational quantum mechanics methods. Providing a shared, Web-accessible environment where researchers can not only download basis sets in their required format but browse the data, contribute new basis sets, and ultimately curate and manage the data as a community will facilitate growth of this resource and encourage sharing both data and knowledge. We describe the Basis Set Exchange (BSE), a Web portal that provides advanced browsing and download capabilities, facilities for contributing basis set data, and an environment that incorporates tools to foster development and interaction of communities. The BSE leverages and enables continued development of the basis set library originally assembled at the Environmental Molecular Sciences Laboratory.

INTRODUCTION

Quantum mechanics and relativity are two of the most important milestones in physics in the 20th century. Quantum mechanics calculations, including those with relativistic corrections, are extensively used in a variety of fields in modern science, including chemistry, physics, material science, biochemistry, and medicinal chemistry, to name a few. The fundamental principles in quantum mechanics are based on the Schrödinger equation[1] or the Dirac−Coulomb[2,3] equation in the case of relativistic quantum mechanics.

These many-electron equations can be approximated by simplifying the problem to the Hartree−Fock method and extended ab initio equations that include electron correlation effects.[4] These equations are usually represented by one-electron wavefunctions, which result in important quantum mechanics concepts like atomic orbitals (AOs) and molecular orbitals. Because the equations cannot be solved analytically via mathematics, the key to solving them is to use iterative numerical computations by fitting the radial part of the wavefunction or electronic density by mathematical functions that are called basis functions or basis sets. These basis functions can be any mathematical functions that form a complete set.

In practice, three kinds of basis functions are widely used in the scientific community because of their computational advantages; they are Slater, Gaussian, and plane-wave. Slater was the first to develop a function called Slater-type orbital or Slater-type functions (STFs).[5] STFs behave similarly to hydrogen-like radial functions and are efficient in representing the AOs, but they are difficult to use in calculating three- and four-center two-electron integrals. Boys[6] started to use Gaussian-type orbitals or more properly Gaussian-type functions (GTFs), which differ from the STFs in the exponent term. When the same number of functions are used in linear least-squares fittings, the GTFs are less accurate than the STFs because they decay too fast when compared with the "real" radial functions (e.g., hydrogen-like orbitals). However, the GTFs are advantageous because the product of two GTFs yields a new GTF, which greatly facilitates computations. The primitive GTFs can be combined together to form basis sets with different levels of accuracy. Commonly used GTFs include minimum (or single-ζ) basis sets, double-ζ, triple-ζ, quadruple-ζ, and so forth, with or without polarization and diffuse functions (for a nice review including the terminology of the Gaussian basis sets, the reader is referred to ref 12). The third widely used basis functions are plane-wave functions, which are advantageous in the treatment of periodic systems.[7,8] Plane-waves have the advantage of simplifying various equations and do not depend on the nuclear positions of the atoms. Their disadvantage is that a large number are usually needed for molecular systems, especially those that are being examined for localized properties. In this paper, we focus on the GTFs because they are the most widely used in the computational chemistry community.

Traditionally, basis sets are developed by relatively few experts and reused widely by the research community. Initial efforts to disseminate basis sets resulted in the compilation of basis set books[9−11] and review articles,[12−14] which required users to manually transcribe the data into their input files, possibly introducing errors, and requiring many individuals to track corrections to the published data. All major codes include an internal library of basis sets which addressed this problem to some extent, however at the cost of creating many basis sets databases with no mechanism to keep them synchronized. With the availability of the Internet, the Web-based Environmental Molecular Sciences Laboratory (EMSL) Gaussian Basis Set Order Form (GBSOF) was developed by Feller,[15] providing a single access point to a growing number of basis sets and offering the advantages of eliminat-

* Corresponding author e-mail: Karen.Schuchardt@pnl.gov.
† Pacific Northwest National Laboratory.
‡ Iowa State University.

10.1021/ci600510j CCC: $37.00
Published on Web 04/12/2007

ing transcription errors while providing the data in a number of common formats, a single source of current "best-available" data, expert validation of data prior to public availability, error reporting, and some basic browsing capabilities. The GBSOF remains very popular; however, maintenance of the data and software relied on the extraordinary efforts of a few people at a single institution with maintenance decreasing over time with staff turnover and changes in priorities. Additionally, the existing capability uses the same technology used in 1994, namely, a set of flat files with custom FORTRAN programs to parse and format the outputs, making it difficult for others to extend and maintain. Finally, Web advances such as Dynamic HTML[16] and servlet[17] technology create the possibility of a much richer interface.

In this paper, we present the Basis Set Exchange (BSE), which combines the data and concepts developed for the GBSOF with a richer Web interface to provide sophisticated mechanisms to browse the data, contribute online, and programmatically access the data. Furthermore, the BSE architecture opens the door to community features including cross-organizational curation, annotations to share knowledge (not just data) among the user community, notifications for data updates, and statistics on data usage patterns.

METHODS

The Basis Set Exchange concept arose from efforts on the Collaboratory for Multiscale Chemical Science (CMCS)[18] project to develop and deploy a Web portal[19] and infrastucture to pilot the concept of a "knowledge grid" for multiscale informatics-based chemistry research. The goal of CMCS was to incorporate advances in informatics, the semantic Web,[20] and collaboratories to facilitate collaboration among and within various subdisciplines of combustion research. Under the "knowledge grid" concept, these communities share data and analysis tools as they create verified, documented data sets and reference data. There are many challenges to realizing this vision: data ownership/licensing issues, developing agreement on what constitutes validated, curated data within a given community, and the lack of reasonably well-vetted data sets to serve as a starting point in developing the tools that empower the community. The EMSL Gaussian Basis Set library was chosen as a pilot data set because there is an existing well-vetted data set already in use by the community, the data was no longer being frequently maintained, the data was available to members of CMCS, and there is a relatively small community of basis set developers with a track record of contributing data. In the following four subsections, we describe the software architecture, the schema used, the data and metadata management, and the Web services available in BSE.

Architecture. The core CMCS architecture,[21] as shown in Figure 1, is comprised of three major components: a community portal, a rich content management system, and a messaging/notification subsystem. Because the capabilities of this core architecture are general-purpose and useful to any community that has the need to develop, publish, and share data and to provide tools to distributed members, it was released as the open-source Knowledge Environment for Collaborative Science (KnECS) toolkit.[22] Both the Basis Set Exchange and the CMCS are extensions of this toolkit. We briefly describe this architecture and the capabilities most relevant to the BSE.

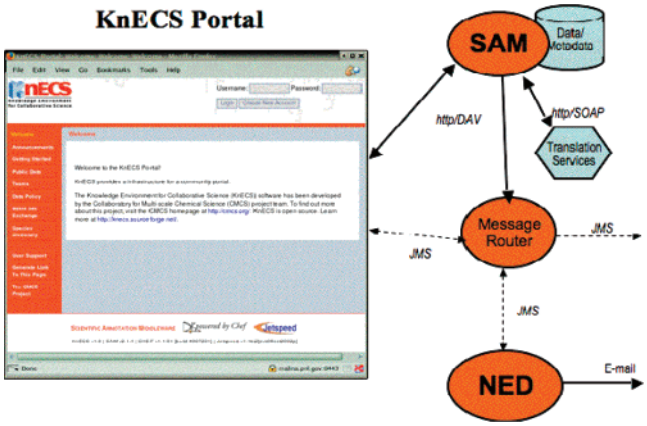

Figure 1. CMCS/KnECS architecture.

The portal is the central integrating component of the KnECS toolkit. A portal provides a Web environment where users can create customized workspaces, access data and applications, and interact with other users. A portal can be extended with "portlets" or custom applications that are then available to users and communities. The BSE itself is an example of a portlet extension to KnECS. The KnECS portal comes with a number of useful capabilities and portlets to facilitate data collaboration. For example, when a user signs up for an account, a private space on the KnECS document repository is created. Users can create teams, where each team is given both a public and private workspace where data can be uploaded and shared. Using the Data Browser portlet, team members can upload and download data, search the repository, control access on a per-resource basis, and request notifications related to access or modification of data under collaborative development. The portal environment also includes team collaboration tools such as a team mailing list, chat, and a task list.

The core of the informatics infrastructure is based on the Scientific Annotation Middleware (SAM) software, which provides scientific content management (data and metadata[23] management).[24] SAM provides a range of capabilities for storing and retrieving data and metadata, searching, versioning, locking, access control, and managing provenance[25] and other data relationships. Of particular relevance is the mechanism to register metadata extractors or translators based on MIME type.[26] The former extract searchable, viewable metadata whenever a document is uploaded, ensuring that any revisions to documents will automatically result in updated metadata, while the latter provide custom, on-demand views of the data to users.

KnECS includes a Java Messaging System (JMS)[27] component to share events. A publish/subscribe messaging system, such as JMS, is a key construct for building loosely coupled systems and provides an extension point for developers and end users. In KnECS, messages are generated anytime any activity occurs on the data server. The main recipient of the messages is the notification e-mail daemon, which compares events to user-created notification requests to create immediate, daily, and monthly digests of activities that can be sent to individual users or entire teams. In the context of the BSE, this capability is used to generate

notifications that new data have been uploaded and could be used to notify users of corrections to data sets.

KnECS has many extension points allowing communities to customize the portal. SAM can be configured to extract custom metadata or provide dynamic views of data files, allowing each community to define and evolve their set of metadata and data formats. Teams can create custom "advanced search" forms simply by creating an XML Schema[28] descriptor file defining the labels and types of data available and uploading this descriptor file. They can also create custom public browsing and views of their data by creating HTML pages with the embedded SEARCH element. In summary, KnECS, and therefore BSE, provides core capabilities as well as extension mechanisms to support the formation and evolution of communities interested in developing and sharing tools and data sets.

Schema. We chose XML[29] as the storage format for basis set data and developed XML Schema definitions which define and enumerate acceptable XML representations of the data. Each basis set is a collection of one or more files managed by the SAM content repository. The two primary XML Schema definition files are (1) Gaussian basis set [including the special case of the density functional theory (DFT) fitting basis sets[30]] and (2) effective core potential (ECP).[31] The schema definitions correlate strongly with the original text-based format developed for GBSOF but are predominantly self-describing and leverage the parsing and data manipulation tools available for XML data. A detailed description of the schemas is beyond the scope of this paper so we limit our discussion to some of the key design concepts. However, a complete listing for each of these schema definition files can be found with this article's Supporting Information. The schema definitions can also be accessed from the portal's *About* page.[32]

One of the strengths of XML Schema is the capability to easily leverage existing "standard" schema definitions through the use of the import mechanism. Schemas were imported from the following resources: Dublin Core,[33] providing standards for cross-domain information resource descriptions; XML Linking Language[34] (XLink), providing capabilities to create and describe link relationships between XML resources; and Chemical Markup Language[35] (CML), providing definitions for common molecular concepts and advanced data types. One example is the use of CML's matrix element. For a specific chemical element and shell type, the matrix element is used to represent the exponents and coefficients (i.e., rows and columns) for a basis set contraction. While this is not fully self-describing, it was adopted for legacy reasons. The following is the portion of the schema definition file that defines the structure for a basis set contraction, along with example data for a contraction taken from the 6-31G[36] basis set document:

```
…
<xs:element name="contraction" maxOccurs="unbounded">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="cml:matrix"></xs:element>
        </xs:sequence>
        <xs:attribute name="shell" type="shellType"/>
    </xs:complexType>
</xs:element>
…

…
<contraction shell="S">
  <cml:matrix rows="3" columns="2" dataType="xs:double">
    18.731137 0.03349460
    2.8253937 0.23472695
    0.6401217 0.81375733
  </cml:matrix>
</contraction>
…
```

Gaussian-type basis sets are represented by exponents and linear least-square coefficients. The ECPs are also expanded in the GTFs. For GTFs, there are typically two ways to represent them, general or segmented contractions.[37] In a general contraction, the primitive GTFs are allowed to contribute to several basis functions. In a segmented contraction, each set of primitive GTFs usually contributes to only one basis function. The basis sets can also use either spherical or Cartesian Gaussian primitives. The BSE schema was designed to ensure that this and other important metadata would be accurately represented.

The schema design also had to take into account that different basis sets may use a common set of GTFs. For example, the 6-31G* basis set is composed of the 6-31G orbital basis set data along with an additional set of polarization functions. The schema could have been designed so that all GTFs for a basis set are included in a single XML document. This, of course, implies duplication of data across XML documents and creates data maintenance issues. This type of data redundancy has been avoided by designing the schema so that a basis set document may describe its relationship to other documents contained in the data repository through the use of XLink. In addition to shared GTFs, XLinks are also used to link a basis set document to constituent documents. For example, in ECP basis sets, the data set consists of the Gaussian basis set linked to its associated ECP data. Note that mulitple Gaussian basis sets may link to the same ECP basis set. Similarily, with DFT basis sets, the orbital basis set is augmented with charge and/or exchange fitting data which adheres to the same schema as the Gaussian basis sets. Finally, for data uploaded through the portal, XLinks are used to refer to supplemental upload data such as references and energies (if supplied).

Data and Metadata Management. The BSE's initial data repository was taken directly from the GBSOF data set. Data files from the GBSOF were first run through a conversion process, which produced a set of XML documents conforming to BSE's schema. These XML documents constitute what is called the baseline *EMSL Library* portion of BSE's data repository. The *EMSL Library* is considered valid; that is, no review or curation is required because this set of data has already gone through such a process. Once conversion was complete, the resulting XML documents were uploaded to the data server.

The BSE utilizes metadata to search, query, browse, and perform curation on the basis set data. Some metadata is key to the description of the data itself and is included in the data files. An example is the contraction type described previously. A second type of metadata is implicit metadata such as the elements supported by the data set. A third type of metadata is "external" metadata such as its current curation status and when it was last changed. For the first two types of metadata, the BSE takes advantage of SAM's automatic metadata generation capabilities by using an XSLT[38] script, registered against the document's MIME type. XSLT metadata extraction scripts were developed for both Gaussian

Table 1. Metadata Used by the BSE

	Dublin Core Definitions
Title	name of the basis set
format	MIME type of the document
Description	detailed description of the basis set
Abstract	refinement of the Description element and used to provide a condensed version or brief description of the basis set
	Basis Set Definitions
basisSetType	identifies the type of basis set (e.g., ECP Orbital)
contractionType	indicates the basis sets contraction scheme, which may be "general", "segmented", or "uncontracted"
hasElements	contains a list of elements covered by the basis set
	Curation Definitions
curator	name of the individual performing the data curation
curatorAffiliation	institution the curator is affiliated with
curationDate	date curation was completed
curationNotes	contains notes or comments made by the curator
curationStatus	indicates status of curation, whether or not the basis set data have been newly contributed, verified, published, unverified, or rejected
contributionType	indicates the type of basis set being contributed; for example, it may be a new basis set or a modification to an existing basis set
contributionNotes	contains notes or comments provided by the contributor
contributorName	name of the person contributing the basis set data
contributorEmail	e-mail of user contributing the basis set data
contributorId	login ID of user contributing the basis set data
contributorPI	name of the primary basis set developer or team lead for a basis set development team
contributorPIEmail	e-mail of the primary basis set developer
	Basis Set Relationship Definitions
primaryBasisSetLink	describes the relationship between the document for which this property exists and another document containing the basis set's primary set of basis set functions; this property only applies to basis sets that consist of more than one basis set document
basisSetLink	describes the relationship between the document for which this property exists and another document containing the basis set's secondary set of basis set functions; this property only applies to basis sets that consist of more than one basis set document; this could be, for example, polarization functions for the 6-31G* basis set
referencesLink	describes the relationship between the document for which this property exists and another document containing the publication reference information for the basis set
effectivePotentialsLink	describes the relationship between the document for which this property exists and another document containing the effective core potential data for the basis set
dftFittingLink	describes the relationship between the document for which this property exists and another document containing the DFT fitting data for the basis set

orbital and ECP schema definitions. The BSE leverages the Dublin Core's Metadata Element Set[39] and Dublin Core's Metadata Terms.[40] It also has access to standard data server properties (size, creation date, etc.) provided by SAM. The metadata properties defined and used by BSE, not including some common file properties such as size, creation date, modification date, and so forth, are shown in Table 1.

The BSE data repository contains the following six distinct types of documents and uses the relationship properties defined in Table 1 to properly associate these documents:

(1) Basis set document type: contains the actual basis set exponent and coefficient data as well as associated basis set metadata

(2) Aggregation document type: does not contain any basis set function data, but instead contains a description of the basis set along with links to the other documents that make up this basis set; this structure accounts for the composition of basis sets described in the previous section

(3) Reference document type: contains publication reference information for the basis set and/or its effective core potentials

(4) Effective core potentials document type: contains the effective core potentials data (scalar or spin−orbit) for a basis set

(5) DFT fitting document type: contains the associated charge and/or exchange fitting basis set

(6) Energies document type (optional): contains energy values to be used for data verification. Applies only to newly contributed data

Figure 2 shows an example of the documents and relationships used to represent the 6-31G* basis set in BSE's data repository.

Web Services. The main capabilities of downloading and contributing basis sets are implemented as Web services enabling some important usage scenarios: dynamic access to basis sets by environments such as the Extensible Computational Chemistry Environment,[41] GridChem,[42] or the E-Science effort in the United Kingdom,[43] and bulk downloads of the latest and best values for inclusion with the release of a variety of computational codes. By providing these capabilities, we are encouraging the development of a single comprehensive community database since both users

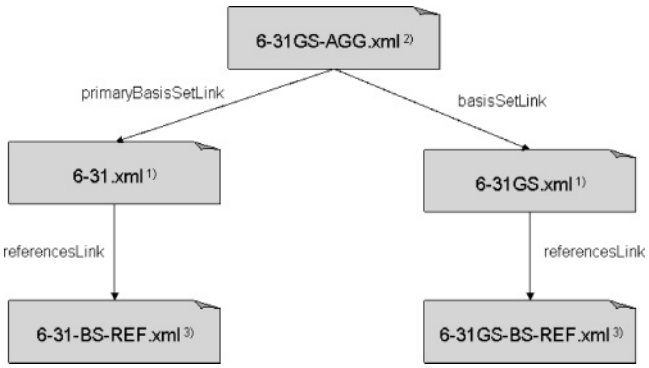

Figure 2. Documents and interdocument relationships used to describe the 6-31G* basis set.

and software tools have ready access to the latest, best data at any time. The Web service interfaces are documented in the interface definition document at http://purl.oclc.org/net/emsl/bse/services. As an example, the methods support querying for the names of available basis sets and then requesting individual basis sets by name. We plan to add additional Web service methods in the future such as filtering the data on the basis of criteria like elements, type, or author.

RESULTS

Using the technologies and techniques described in the previous section, we have developed and deployed a Web-based environment (http://purl.oclc.org/net/emsl/bse) for downloading and uploading GTF basis sets, including all-electron basis sets, ECPs, and ECP valence basis sets. The Web interface supports three types of users: anonymous users with access to verified or published data from the download page, contributors who can upload their data sets, and curators who are responsible for verifying contributed data and performing data maintenance. All users have access to the support queue, which is directly integrated with EMSL's Molecular Science Computing Facility support software. Only curators have access to the full set of tools provided by the KnECS infrastructure. The following sections describe the key capabilities of the user interface.

Browse and Download. The download page, shown in Figure 3, supports multiple methods for browsing the library, in addition to supporting the download function. The list box on the left contains all published or verified basis sets and can be filtered by type—currently All Electron Orbitals, ECPs, or all. Contributors and curators can also filter on status and therefore see new or unverified basis sets. This capability essentially allows basis set developers to use the BSE as both a reference and a development tool. When a basis set is selected, an orange triangle in the lower left corner of relevant periodic table buttons shows which elements are included in the basis set. Information about the selected basis set shows up in the bottom portion of the page, and more detailed data, including references and developer comments, is available through the "more information" link.

Selection of a particular element in the periodic table (shown as green in Figure 3) serves two purposes: to specify what element data should be included in a download request and to filter the list of basis sets to only show those that have data for all selected elements. For example, in the figure, basis set cc-pwCV5z includes data for elements B−Ne and Al−Ar, but only boron will be included in the download (since it alone is selected). The list of basis sets has been reduced to the 131 sets that have data for this element. Through these mechanisms, the download page provides a detailed at-a-glance view of individual basis sets and enables a user to quickly filter and browse through the entire library.

The download function currently supports 15 of the most commonly used formats from the original GBSOF and provides more consistent formatting of downloaded data so it can be inserted easily into computational input files.

Contribute. Contributors are required to create an account with the BSE, which provides both a means to contact the contributor if an inconsistency arises during validation and also reduces the risk of deliberate attacks. When a contributor logs in, the menu bar will include an additional "Contribute" menu item. The contribution process is performed by following a series of steps. The initial step is informational and provides an overview of the full process. During the second step, one of four purposes must be chosen: contribute a new basis set, contribute an orbital data set and associate it with an existing ECP, add new elements to an existing basis set, or modify/correct an existing basis set. On the basis of the choice, the interface will present slightly different fields. Figure 4 shows the fields required to contribute a new basis set. The contributor provides a proposed name and selects one of the four supported import formats: NWChem, Gaussian, GAMESS (US), and Molpro. The contributor also specifies the harmonic type, the local path to their file, and a description of the basis set. The primary developer and e-mail fields allow multiple members of development groups to upload data while maintaining an association to the research lead.

Next, references can be provided, although they are optional. Entry of multiple references is supported, and each reference can be associated with a set of specified elements. The next step allows the input of Hartree−Fock energy data for each element for validation purposes. The contributor must select the method of input for their energy data by either filling in a provided table, specifying a file containing the data, or choosing "Not available". Input of energy values is optional, but we require either reference information or energy values for a basis set to be made publicly available for download. After all data has been entered, a review page allows the user to verify the data before completing the process. During this process, the basis set data is parsed and converted to the BSE schema and then translated back to the original format essentially providing a round-trip test of our parser and translator and enabling the contributor to verify that their data will be handled correctly. The contributor may also provide a note to the curator to indicate any extra information that will assist the curator in verifying or integrating the data set. An example of the latter would be information on the relationship of the newly contributed data to other existing basis sets and how that might affect curator actions (replace data set or create a new one).

When the contribution process completes, the new data is given a status of "New" and must be examined, verified, and possibly edited by a curator before becoming generally available. If the contributor or another individual finds an error, the appropriate action would be to use the "Help"

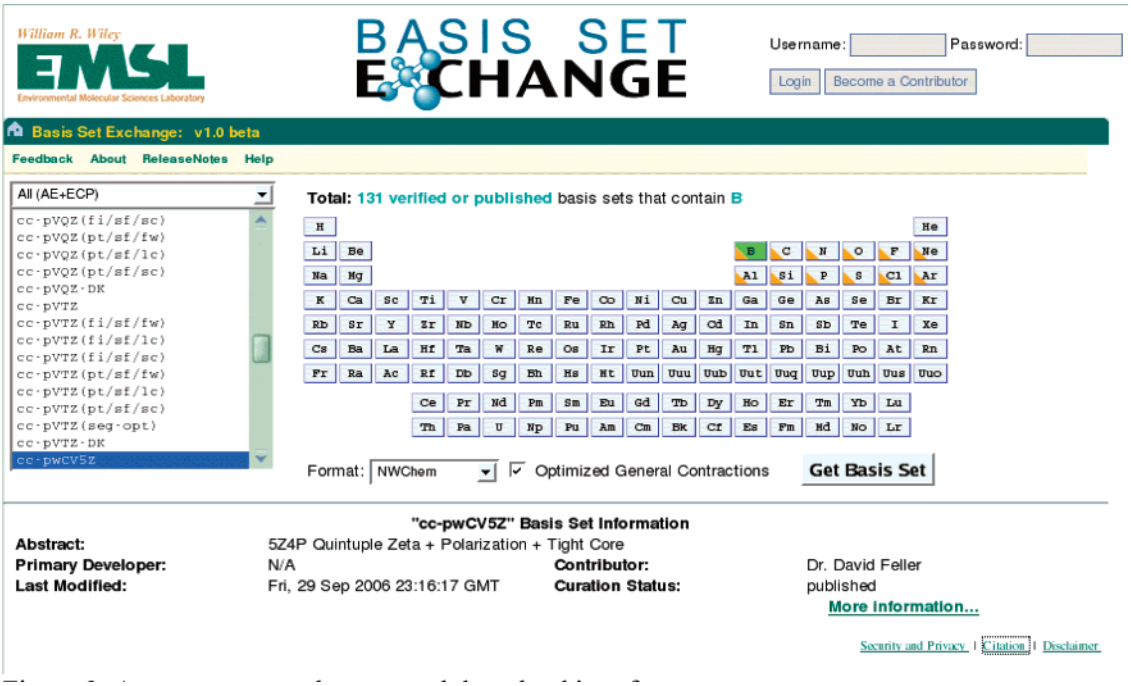

Figure 3. Anonymous user browse and download interface.

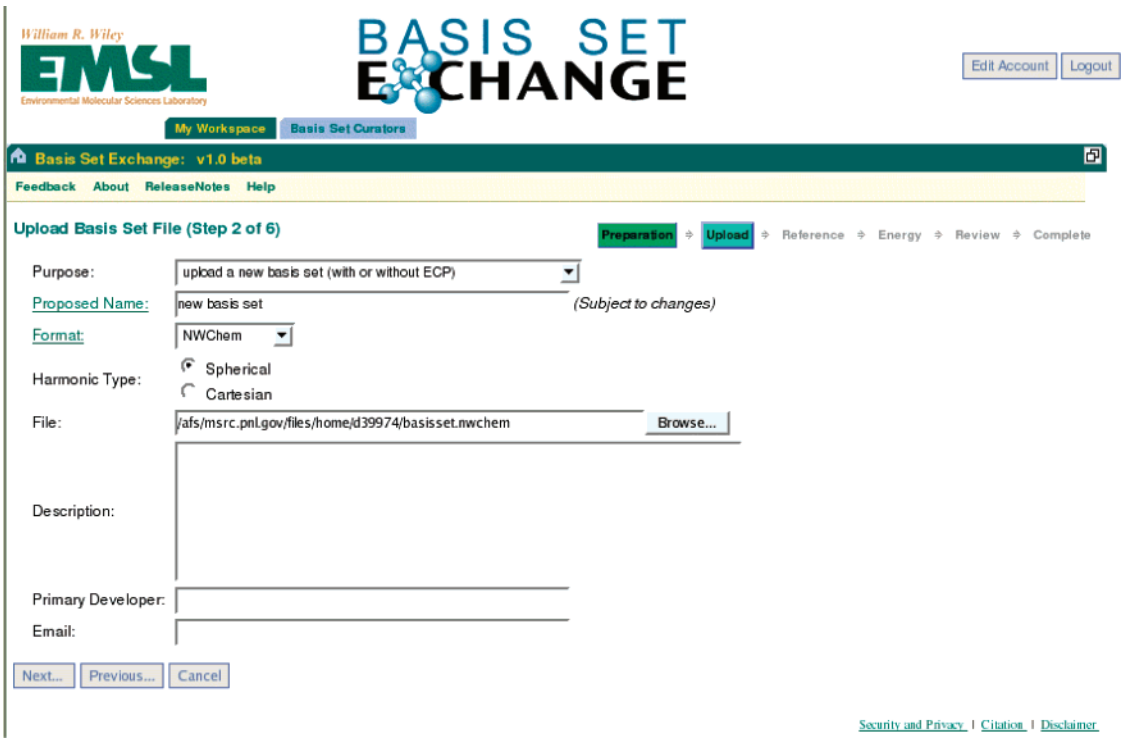

Figure 4. Upload page.

button to send a queue request for a curator to address the problem.

Curation. The EMSL Basis Set Library and associated GBSOF have been successful in part because the data disseminated by this service have proven to be very accurate. Acquiring quality basis set data is a significant time-saver for this community. This accuracy was achieved because of individuals within EMSL taking ownership of the required data curation activities. BSE intends to more formally support these curation activities within a distributed team environment and develop additional tools to simplify these tasks over time. Curation activities include reviewing new submissions, performing verification calculations when energies are available, verifying references, making data corrections, responding to queue items, integrating basis set extensions, assigning curation status, and adding annotations to the data. Community tools provided through the KnECS infrastructure support these activities in several ways. First, notifications have been configured to send e-mail to the curation team any time data is uploaded—freeing curators from manually monitoring the system for new submissions. The data browser tool provides a file system type view of the basis set files that curators can view, download, edit, and upload as revised versions of the data as needed. The data browser

also contains a BSE form for modifying the curation status as consensus is reached on the data. A custom, advanced search form was developed by providing an XML configuration file describing the most important search fields from the list of metadata elements described previously. This allows curators to quickly search for basis set data on the basis of these particular metadata fields. Additional search forms and search fields can be easily added and/or modified on the basis of curator preferences.

Initial curation responsibilities will be carried out by EMSL staff. However, our goal is to open this process to other community members with expertise and interest in maintaining and growing the database. The BSE does not reduce the overall effort but allows it to be distributed among a wider group and ameliorates the cost to a single institution. As this transition occurs, other tools such as the team mailing list and team task list will help to coordinate activities.

DISCUSSION

In this work, we have presented the BSE, a Web-based environment that allows researchers to download, upload, and curate Gaussian basis sets. The architecture is flexible and will accommodate modifications and extensions as the community evolves its use of the environment. Since the data representation is based on standards such as XML and XML Schema, and is openly accessed through Web services, the data is now accessible to applications as well as users. We anticipate that these features will make the BSE a more valuable community resource than the original GBSOF, which has proven to be an essential resource for the computational science community.

The choice of XML for the data format enabled the use of standard technologies such as XSLT to perform metadata extraction and format conversions. In practice, we found XSLT to be well-suited to metadata extraction but insufficient for format conversion for a number of reasons. Formatting numeric data in aligned columns for human consumption, while possible, was very tedious and required an extensive amount of XSLT. Additionally, while many of the code formats are organized similarly to the XML element structure, other formats are organized in such a way that requires extensive looping, a task not explicitly supported by XSLT. Finally, the conversion process is somewhat more complex than simple text conversion. For codes that do not support the generally contracted syntax, the data must be reformatted into segmented notation. Also, with the "Optimize General Contraction" option, additional format conversion must take place, and the logic for doing this depends in part on evaluating the numeric values of the data in the matrix. This might be somewhat easier to do with schema improvements such as replacing the cml:matrix element with fully self-describing constituent elements. Though these obstacles can be overcome, concern for maintenance led us to switch to Java-based conversion scripts. The use of XML still delivered useful tools such as validating parsers and standard application programmer interfaces. Writing the converters in Java required substantially less effort that XSLT.

There are a number of enhancements to the BSE that could make it an even more valuable resource. Providing access to the data in its native XML format would facilitate code interoperability. Bulk access to the native XML would also allow computational codes to readily synchronize with the latest reference data set, solidifying the BSE as the single, community-controlled repository of basis sets. Providing these capabilities would likely require some iteration on the schemas by the community, as the initial schema was developed for the purpose of describing a reference data set adhering closely to the legacy format. Tools to better support data curation and knowledge sharing are needed. In particular, tools to assist in editing and merging corrections would be of significant benefit. Tools to annotate basis sets are desired to support our vision of capturing community knowledge and experience. Along these same lines, incorporating atomic energies in a manner that they can be readily queried and updated would further build upon the knowledge base. Portal tools such as visual notification of updated or new data, along with access statistics gathering and reporting would benefit basis set contributors as well as users. Finally, there are additional download code formats we hope to support, and we are also interested in extending these concepts to support other types of basis functions such as plane-wave and Slater.

While BSE does not reduce the effort involved in curating basis sets, it does open the process to the community of researchers and users that benefit from the data. It is hoped that, by shifting the burden from a few people at a single organization to the broader community, the data set will expand more quickly and, over time, more knowledge about the data will be shared. It is further hoped that the BSE can serve as a starting point for bringing the power of the participatory Web to similar scientific data sets.

ACKNOWLEDGMENT

The authors wish to acknowledge the extensive efforts of Dr. David Feller in establishing the original EMSL Gaussian Basis Set Library that has been a valuable resource to the computational chemistry research community for many years. The BSE software was developed by the CMCS project and has been further customized and deployed with financial support through EMSL as funded by the Office of Biological and Environmental Research in the U.S. Department of Energy (DOE). PNNL is operated by Battelle for the U.S. Department of Energy under contract DE-AC06-76RLO 1830. Funding for the CMCS project was provided by the Mathematics, Information and Computer Science Division of DOE through the Scientific Discovery through Advanced Computing program.

Supporting Information Available: The two primary XML Schema definition files, which enumerate acceptable representations of the basis set data, are (1) Gaussian basis set (including the special case of the DFT fitting basis sets[30]) and (2) ECP.[31] This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.* **1926**, *28*, 1049−1070.
(2) Dirac, P. A. M. The Quantum Theory of the Electron. *Proc. R. Soc. London, Ser. A* **1928**, *117*, 610−624.
(3) Dirac, P. A. M. A Theory of Electrons and Protons. *Proc. R. Soc. London, Ser. A* **1930**, *126*, 360−365.
(4) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, 2nd ed.; MacMillan Publishing Co.: New York, 1989.

(5) Slater, J. C. Analytic Atomic Wave Functions. *Phys. Rev.* **1932**, *42*, 33−43.
(6) Boys, S.F. Electronic Wave Functions I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc. R. Soc. London, Ser. A* **1950**, *200*, 542−554.
(7) Herring, C. A New Method for Calculating Wave Functions in Crystals. *Phys. Rev.* **1940**, *57*, 1169−1177.
(8) Phillips, J. C.; Kleinman, L. New Method for Calculating Wave Functions in Crystals and Molecules. *Phys. Rev.* **1959**, *116*, 287−294.
(9) Poirier, R.; Kari, R.; Csizmadia, I. G. *Handbook of Gaussian Basis Sets*, 1st ed.; Elsevier Science: New York, 1985.
(10) Andzelm, J.; Klobukowski, M.; Radzio-Andzelm, E.; Saski, Y.; Tatewaki, H. Gaussian Basis Sets for Molecular Calculations, 1st ed.; Huzinaga, S., Ed.; Elsevier: Amsterdam, The Netherlands, 1984.
(11) Dunning, T. H.; Hay, P. J. Gaussian Basis Sets for Molecular Calculations. In *Modern Theoretical Chemistry*, 1st ed.; Schaefer, H. F., III, Ed.; Plenum Press: New York, 1977; Vol. 3, pp 1−28.
(12) Feller, D.; Davidson, E. R. Basis Sets for Ab Initio Molecular Orbital Calculations and Intermolecular Interactions. In *Reviews in Computational Chemistry*, 1st ed.; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1990; pp 1−43.
(13) Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(14) Ahlrichs, R.; Taylor, P. R. The Choice of Gaussian Basis Sets for Molecular Electronic Structure Calculations. *J. Chim. Phys.* **1981**, *78*, 315−324.
(15) Feller, D. EMSL Gaussian Basis Set Order Form. http://www.emsl.pnl.gov/forms/basisform.html (accessed Jan 12, 2007).
(16) Dynamic HTML Overview. http://en.widipedia.org/wiki/Dynamic_html (accessed Jan 22, 2007).
(17) Java Servlet Technology Overview. http://java.sun.com/products/servlet/overview.html (accessed Jan 18, 2007).
(18) Collaboratory for Multi-Scale Chemical Science. http://cmcs.org/ (accessed Jan 12, 2007).
(19) Web Portal Overview. http://en.wikipedia.org/wiki/Web_portal (accessed Jan 22, 2007).
(20) Semantic Web. http://www.w3.org/2001/sw/ (accessed Jan 12, 2007).
(21) Myers, J. D.; Allison, T. C.; Bittner, S. J.; Didier, B. T.; Frenklach, M.; Green, W. H.; Ho, Y.; Hewson, J.; Koegler, W. S.; Lansing, C. S.; Leahy, D.; Lee, M.; McCoy, R.; Minkoff, M.; Nijsure, S.; von Laszewski, G.; Montoya, D.; Oluwole, L.; Pancerella, C. M.; Pinzon, R.; Pitz, W.; Rahn, L. A.; Ruscic, B.; Schuchardt, K. L.; Stephan, E. G.; Wagner, A.; Windus, T. L.; Yang, C. A Collaborative Informatics Infrastructure for Multi-Scale Science. *Cluster Comput*. **2005**, *8*, 243−253.
(22) Schuchardt K. L., Didier, B. T.; Kodeboyina, D.; Leahy, D.; Myers, J. D.; Oluwole, O.; Pancerella, C. M.; Pitz, W.; Rahn, L.; Ruscic, B.; Song, J.; Laszewski, G. V.; Yang, C. Portal-Based Knowledge Environment for Collaborative Science. *Concurrency Comput.: Practice Experience* in press.
(23) Metadata Overview. http://en.wikipedia.org/wiki/Metadata (accessed Jan 22, 2007).
(24) Myers, J. D.; Chappell, A.; Elder, M.; Geist, A.; Schwidder, J. Re-Integrating the Research Record. *IEEE Comput. Sci. Eng. 2003* **2003**, *5*, 44−50.
(25) Provenance Overview. http://en.wikipedia.org/wiki/Provenance (accessed Jan 22, 2007).
(26) Mime type definition, from the World Wide Web Consortium Web site. http://www.w3.org/2003/01/xhtml-mimetype/ (accessed Jan 12, 2007).
(27) Sun Messaging Systems and the Java Message Service (JMS). http://java.sun.com/developer/technicalArticles/Networking/messaging/ (accessed Jan 12, 2007).
(28) Requirements for XML Schema 1.1, January 2003, from World Wide Web Consortium Web site. http://www.w3.org/TR/xmlschema-11-req/ (accessed Jan 12, 2007).
(29) Extensible Markup Language (XML) 1.0 (Fourth Edition), September 2006, from World Wide Web Consortium Web site. http://www.w3.org/TR/xml/ (accessed Jan 12, 2007).
(30) See, for example, Dunlap, B. I. Fitting the Coulomb Potential Variationally in *X*α Molecular Calculations. *J. Chem. Phys.* **1983**, *78*, 3140−3142.
(31) See, for example, Krauss, M.; Stevens, W. J. Effective Potentials in Molecular Quantum Chemistry. *Annu. Rev. Phys. Chem.* **1984**, *35*, 357−385.
(32) Basis Set Exchange Schema Definition. http://purl.oclc.org/net/emsl/bse/docs/schemas/ (accessed Jan 22, 2007).
(33) Dublin Core Metadata Initiative Web site. http://dublincore.org/ (accessed Jan 12, 2007).
(34) XML Linking Language (XLink) Version 1.0, June 2001, from World Wide Web Consortium Web site. http://www.w3.org/TR/xlink/ (accessed Jan 12, 2007).
(35) Murray-Rust, P.; Rzepa, H. Chemical Markup Language (CML) Web site. http://www.xml-cml.org (accessed Jan 12, 2007).
(36) **(36)** Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724−728.
(37) Raffenetti, R. C. General Contraction of Gaussian Atomic Orbitals: Core, Valence, Polarization and Diffuse Basis Sets; Molecular Integral Evaluation. *J. Chem. Phys.* **1973**, *58*, 4452−4458.
(38) XSL Transformations (XSLT) Version 1.0, November 1999 from World Wide Web Consortium Web site. http://www.w3.org/TR/xslt (accessed Jan 12, 2007).
(39) Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2003, from Dublin Core Metadata Initiative Web site. http://www.dublincore.org/documents/dces/ (accessed Jan 12, 2007).
(40) Dublin Core Metadata Terms, 2003, from Dublin Core Metadata Initiative Web site. http://dublincore.org/documents/2003/03/04/dcmi-terms/ (accessed Jan 12, 2007).
(41) Black, G. D.; Schuchardt, K. L.; Palmer, B. J.; Gracio, D. K. The Extensible Computational Chemistry Environment: A Problem Solving Environment for High Performance Theoretical Chemistry; International Conference on Computational Science 2003, Saint Petersburg, Russian Federation, June 2−4, 2003.
(42) GridChem. Computational Chemistry Grid. https://www.gridchem.org/ (accessed Jan 12, 2007).
(43) National e-Science Centre. http://www.nesc.ac.uk (accessed Jan 12, 2007).

CI600510J