ThermInfo: Collecting and Presenting Thermochemical Properties

Ana L. Teixeira[1*], Rui C. Santos[2], Francisco M. Couto[1]

[1] *LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016, Lisboa, Portugal*
[2] *Centro de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal*

Abstract. Due to the large amount of chemical data available, it is essential to organize them and given the numerous problems derived from applying spreadsheets, it motivated us to construct an improved tool geared to contain large amounts of data.

This paper presents *ThermInfo*, a public web tool that integrates a database based upon a relational data specification for describing structural and thermochemical properties of organic compounds. Its user-friendly web interface allows a text-based searching, compounds insertion, and data management.

At present time, *ThermInfo* contains critically evaluated and thermodynamically consistent experimental thermochemical properties values for about 3,000 unique and non redundant compounds. Interface usability results show that it is easy and fast to learn which improves the efficiency when employing the tool.

In the future, we intend to expand the dataset, integrate other classes of compounds databases, use chemoinformatic tools to combine a structure drawn with textual query terms and use prediction methods of thermochemical properties.

ThermInfo is available at http://www.therminfo.com.

Keywords: Chemoinformatics, chemical database, data management, data integration

1 Introduction

In large part, Chemistry is still an empirical science, building its development on an ever increasing amount of data and information. The quantifiable chemo-information keeps growing exponentially due to constantly refined and optimized experimental technologies. According to *Chemical Abstracts Service*, there are currently more than 35 million compounds known, increasing with more than 1 million each year, and more than 600,000 publications per year are related to Chemical information.[1] Thus, it was realized that this flood can only be managed by exploring it in electronic form.

Chemical research work often requires search in many kinds of compounds properties. The problem lies in looking all up in handbooks and papers which is quite unprogressive and time demanding. Thus, it is very important to develop databases to

[*] To whom correspondence should be addressed (ateixeira@lasige.di.fc.ul.pt)

[1] CAS Statistical Summary: [http://www.cas.org/]

manage and search chemical data (such as thermochemical properties) and make them publicly available. Most databases of thermochemical properties are privately owned, expensive, often not so well administrated, and are available through restricted interfaces that are not suitable for the development of statistical and prediction methods [1-3].

The Molecular Energetics Group[2] of Centro de Química e Bioquímica from Faculdade de Ciências, Universidade de Lisboa, has been collecting and critically evaluating experimental thermochemical data of organic compounds from literature. This evaluation is performed through the comparison of values for similar compounds and/or applying additivity methods, and the data is stored in spreadsheets. However, this process triggers several problems, such as:

1. Referential Data Integrity – when it is needed to change or delete some values of a compound property in multiple rows, the same action needs to be repeated several times. Furthermore if some row is forgotten the data will become inconsistent, ambiguous and loses integrity;
2. Data Redundancy – it is directly connected to the previous problem. The data storing is very ineffective due to the same data being entered various times and thereby growing unnecessarily, thus requiring more computer resources (larger size of data and slower access). A good example of this is storing and summarizing the data for the organic compounds characteristics used in *ThermInfo*, since it is a many-many relationship where 28 columns with characteristic names were needed;
3. Data Validity and Non-Uniformity – humans inserting data can be unreliable and there are many ways of entering the same data. The spreadsheets are unable to efficiently identify data errors and different spellings;
4. Limiting Data View – a spreadsheet displays all rows and columns of the table which is bothersome in case of very large datasets, such as *ThermInfo*'s one. The dataset was composed by 3 different sheets, 57 columns and around 3,000 lines. Additionally, there is a lack of detailed sorting and querying abilities. The spreadsheets are designed to analyze data and sort list items, not for long-term storage of raw data;
5. Performance and Capacity – low efficiency with large datasets due to the lack of indexes for accelerating the process of data search, and memory problems due to the load of unnecessary data;
6. Multiuse and Data Entry – data entry forms are not enabled and when the data cannot be conveniently displayed in tabular view, fields for data entry are placed loosely within the spreadsheet. The multiuse requires a lot of discipline, attention, and knowledge from the users which is ineffective and data can be easily corrupted for larger groups of people;
7. Sharing it with the Scientific Community – it is difficult keeping a single file with all the updates and restricting the users from accessing and updating information;

[2] Molecular Energetics Group web page: http://cqb.fc.ul.pt/menergetics/

8. Evolution – the integration of new data, application of prediction methods, and developing other applications that use the dataset is difficult in a spreadsheet format.

All aspects mentioned above motivated us to develop our own system, *ThermInfo*, to collect and present structural information and thermochemical properties, tuned and adapted to meet our specific needs. The goal is to obtain a high quality system that evolves with time, works efficiently according to the purpose, furthermore overcoming the problems of spreadsheets, which guarantees the highest quality, uniqueness and consistency of each entry, plus cost effective to maintain and enhance. Additionally, *ThermInfo* is expected to lower the costs of research and provide an increased number of useful leads, especially when integrated with properties prediction tools.

The rest of this paper is structured as follows: Section 2 describes basic concepts of Chemistry; In Section 3 we present the *ThermInfo* system and its implementation; Section 4 presents the achieved results and discusses them; Finally, in Section 5, we express our main conclusions and directions for future work.

2 Basic Concepts

High-throughput analytical methods for thermochemistry research generate heterogeneous data formats. *ThermInfo* contains experimental data about organic compounds, which can be divided into three categories:

1. *Structural data* – composed by descriptors that specify the molecular structure of the compound [4]:

- Molecular ID, is a unique and stable identifier for the entity with the format CONNNNN (N = digit);
- Compound Name, is the name provided for an entity based on the current recommendations of *International Union of Pure and Applied Chemistry* (IUPAC)[3];
- CAS Registry Number (CASRN), is a unique numerical identifier created and assigned to a chemical substance by *Chemical Abstract Service* (CAS), it does not have any chemical significance and is assigned in sequential order to assure uniqueness. It has the format NNNNNNN-NN-N (1-7 digits, hyphen, 2 digits, hyphen, 1 digit)[4]. The right-most digit is a check digit used to verify the validity and uniqueness of the entire number and it is calculated by taking the last digit times 1, the next digit times 2, the next digit times 3 and so on, adding all these up, and computing the sum modulo 10. For example, the CAS

[3] IUPAC recommendations: http://www.iupac.org/web/ins/2001-043-1-800

[4] CAS Registry Number [http://www.cas.org/expertise/cascontent/registry/regsys.html]

number of methanol is 67-56-1: the checksum 1 is calculated as (6×1 + 5×2 + 7×3 + 6×4) = 61; 61 mod 10 = 1;

- Molecular Formula, identifies each constituent element of a compound by its chemical symbol and indicates the number of atoms of each element in subscript after the chemical symbol. The atoms are in CHXNOS (X = halogen) order. For example, the molecular formula of Benzenemethanol is C_7H_8O, which indicates that it has 7 atoms of carbon, 8 atoms of hydrogen and 1 atom of oxygen;
- Chemical Structure, is a 2D structural diagram of the compound in JPG format;
- Molecular Weight, is the mass of one molecule of that compound, relative to the unified atomic mass unit;
- Physical State, are the distinct forms that different phases of matter take on, and can be: gas, liquid or crystal;
- SMILES (Simplified Molecular Input Line Entry System), is a specification for describing the structure of chemical molecules using short ASCII strings. An important point about this is that there is a difference between upper and lower cases. For example, the cyclohexane has the SMILES C1CCCCC1, while the benzene has the SMILES c1ccccc1 [5];
- USMILES, is a special and unique SMILES among all valid possibilities for a given compound [6];
- Class, Subclass, Family, are hierarchical classifications according with the compound structure;
- Characteristics, are tags according with the functional groups present in the molecule and other characteristics of the compound.

2. *Thermochemical Data* – is related with the energy released or absorbed in chemical reactions, or in physical state transformations [7]:

   - Standard Molar Enthalpy of Formation and the associated error (in $kJ \cdot mol^{-1}$) for:
     - Crystalline Phase;
     - Liquid Phase;
     - Gas Phase;
   - Standard Molar Enthalpy of Phase Change and the associated error (in $kJ \cdot mol^{-1}$);
   - Observations, free text about the enthalpies values.

3. *References Data* – complete references about the compound data, including: author(s), journal/book title, year, volume, and pages.

3 Design Issues and Implementation

To develop *ThermInfo* we are doing multiple interactions of the following steps: system planning; requirements analysis; design; implementation; and maintenance. It is important to note that all these steps are monitored by users, in the first steps to evaluate their own and system needs, and in the last steps, to evaluate the system design and workability [8].

3.1 Functional Requirements

The major features of *ThermInfo* can be divided into two classes, according to its function (represented in Figure 1 [9]):

1. *Search for a Compound*

   ThermInfo has two search methods available to enable both simple and complex queries:

   - *Simple Search*, provides a single text box that allows the users to search for an organic compound in the database, by entering a search term, such as, its name, molecular formula, molecular ID, CASRN, or SMILES;
   - *Structural Search*, provides multiple search fields that allow users to limit the search results to specific compound characteristics.

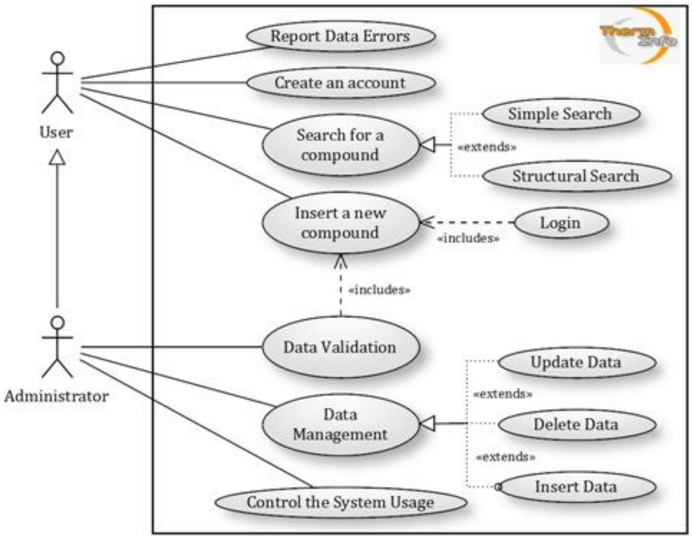


Figure 1 – General overview of the functionality provided by *ThermInfo* in terms of actors, goals and dependencies between use cases.

2. *Data Management*

   2.1 *Insert Data* – allows the users to insert new organic compounds to the database. However, the insertion of the structural, thermochemical, and reference(s) data is restricted to registered users. Before making this new data public an automatic pre-processing stage and a validation process, by an administrator, are performed in order to filter incorrect, ambiguous or non-unique data. The main purpose of this feature is to support the expansion of the database by the scientific community.

2.2 *Report Data Errors* – if users find errors in *ThermInfo* database, they can report it to an administrator by email.

2.3 *Delete Data* – allows the administrator to search for an outdated compound and delete it from the *ThermInfo* database.

2.4 *Update Data* – allows the administrator to search for outdated/erroneous data about a compound and change it.

2.5 *Validate Data* – allows the administrator to check for the new compounds inserted by *ThermInfo* users and approve or reject the insertion of new data into the *ThermInfo* database.

2.6 *Control Panel* – allows the administrator to monitor the usage and the growth of the database.

3.2 *ThermInfo* Database

We have designed a relational database which has to be flexible by eliminating redundancy, inconsistent dependency and to accommodate heterogeneous data: the structural and thermochemical properties of organic compounds and bibliographic references, selected and carefully evaluated from relevant scientific literature and described in the previous section. Figure 2 shows the Unified Modeling Language (UML) class diagram of the *ThermInfo* database, color coded according to the three categories of data and including the entities, relationships and attributes (with data type) [9].

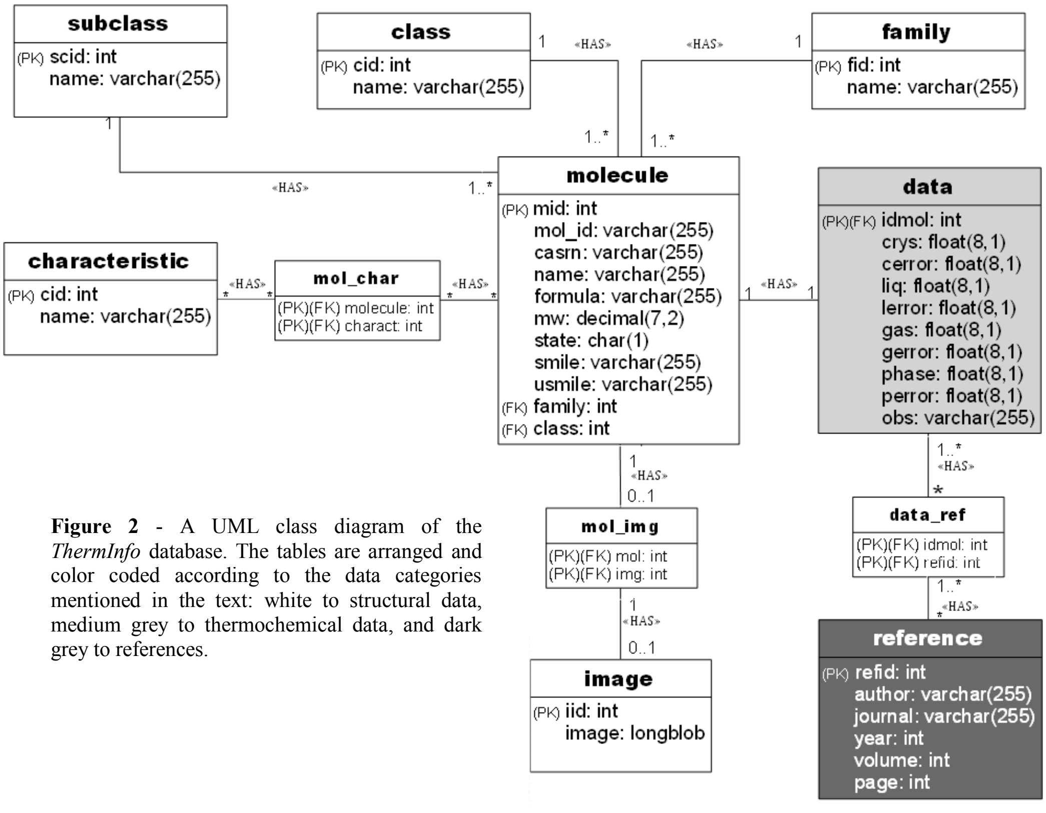


Figure 2 - A UML class diagram of the *ThermInfo* database. The tables are arranged and color coded according to the data categories mentioned in the text: white to structural data, medium grey to thermochemical data, and dark grey to references.

The structural data category is composed by six related entities: the main entity, *molecule*, which contains the data that characterizes the compound and the enforcement of uniqueness assured by the CASRN, since it is a unique descriptor; each molecule can have one *image* with the chemical structure, that is specific to each molecule; each *molecule* can be classified by one *class*, *subclass* and *family*; each *class*, *subclass* and *family* can include several *molecule's;* each molecule can have several *characteristics* or vice versa. The thermochemical properties data category is composed by one entity, *data*, which exists for one molecule (or vice versa). The references category is composed by one entity, *reference*, which can be related to the entity *data*.

3.3 Interface

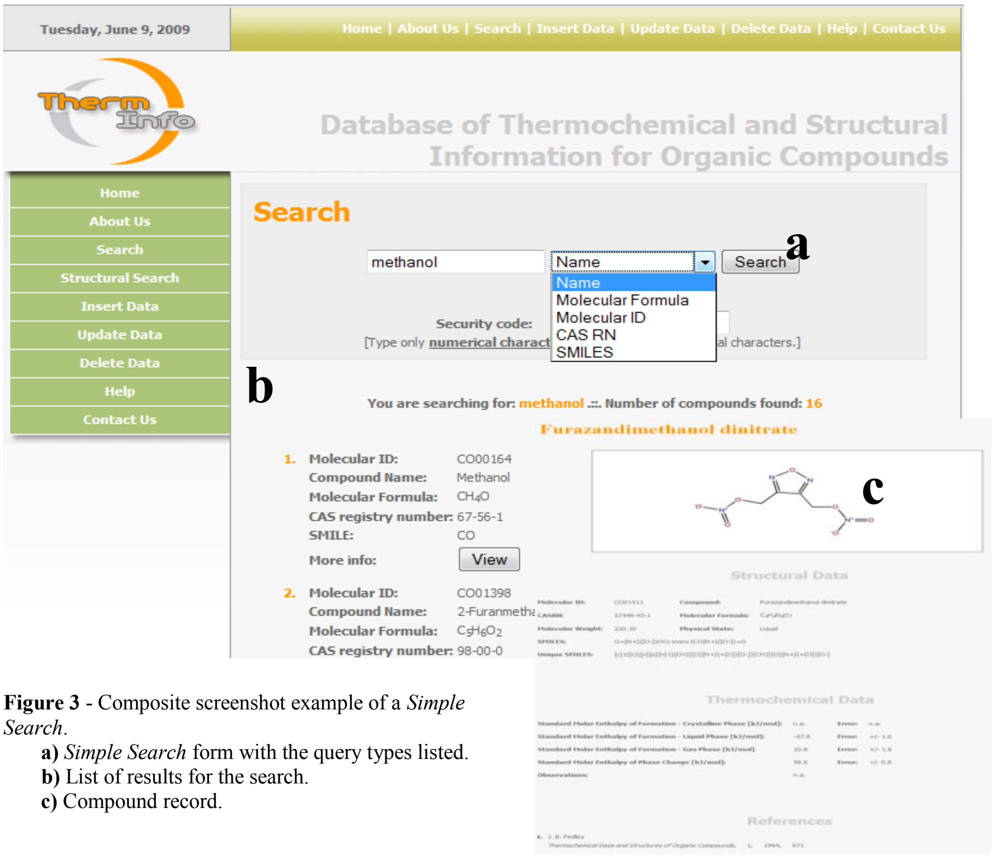
The web interface was developed according to the requirements described in the previous sections. Similar forms were applied to make it user friendly and fast learning. Figure 3 depicts a composite screenshot from the *ThermInfo* search interface. For searching, the user can use the *Simple Search* form (Figure 3a), typing the search term in the indicated format, select one of the five types of search, and type the security code (that protects our database against malicious scripts). If the search term is not in the correct format or the CASRN does not verify the check digit, the users will receive the adequate message, otherwise they will receive a list of maximum 100 most relevant hits, with the molecular ID, compound name, molecular formula, CASRN, and SMILES (Figure 3b), and a link to the complete information record of the compound (Figure 3c[5]). The *Structural Search* (Figure 4) uses the same procedure, but here the user can specify the structural characteristics of the compound and refine the search.

To *Insert* a new compound, the user needs to create an account or login into an existing one. A form with fields for structural, thermochemical, and references data is available. The data inserted will suffer a pre-processing step and the adequate messages will be presented if something is wrong. The inserted data will be placed in temporary tables waiting for an administrative validation.

To *Delete* or *Update* a compound, the administrator needs to search for the compound using its molecular ID. If it exists, it will display the compound information that can be deleted or changed. The outdated data will be moved to outdated tables.

To *Validate* a compound inserted by users, the administrator is able to see the compound data displayed in a table and mark it to be inserted on the database or to be moved to the outdated tables. The decision is automatically reported to the depositor.

[5] This example can be viewed with more detail at the development website http://xldb.di.fc.ul.pt/biotools/therminfo/compound.php?mid=1511&info=View



Tuesday, June 9, 2009

Home | About Us | Search | Insert Data | Update Data | Delete Data | Help | Contact Us
Therm Info
Database of Thermochemical and Structural Information for Organic Compounds
Home
About Us
Search
Structural Search
Insert Data
Update Data
Delete Data
Help
Contact Us
a
b
c

Search

methanol | Name ▾ | Search

Name
Molecular Formula
Molecular ID
CAS RN
SMILES

Security code:
[Type only numerical charact... al characters.]

You are searching for: methanol .::. Number of compounds found: 16

1. Molecular ID: CO00164
Compound Name: Methanol
Molecular Formula: CH_4O
CAS registry number: 67-56-1
SMILE: CO
More info: View

2. Molecular ID: CO01398
Compound Name: 2-Furanmetha
Molecular Formula: $C_5H_6O_2$
CAS registry number: 98-00-0

Furazandimethanol dinitrate

Structural Data

Molecular ID:	CO01511	Compound:	Furazandimethanol dinitrate
CASRN:	[illegible]	Molecular Formula:	$C_4H_4N_4O_7$
Molecular Weight:	220.10	Physical State:	Liquid
SMILES:	[illegible]		
Unique SMILES:	[illegible]		

Thermochemical Data

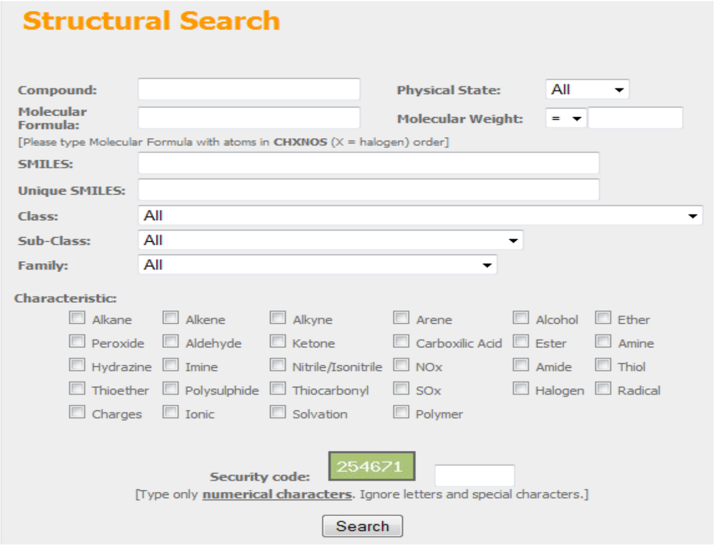
Standard Molar Enthalpy of Formation - Crystalline Phase [kJ/mol]:	n.a.	Error:	n.a.
Standard Molar Enthalpy of Formation - Liquid Phase [kJ/mol]:	-47.8	Error:	+/- 1.6
Standard Molar Enthalpy of Formation - Gas Phase [kJ/mol]	10.8	Error:	+/- 1.8
Standard Molar Enthalpy of Phase Change [kJ/mol]:	58.6	Error:	+/- 0.8
Observations:	n.a.		

References

1. J. B. Pedley
Thermochemical Data and Structures of Organic Compounds, 1, 1994, 571

Figure 3 - Composite screenshot example of a *Simple Search*.

a) *Simple Search* form with the query types listed.
b) List of results for the search.
c) Compound record.



Structural Search

Compound: | Physical State: All ▾
Molecular Formula: | Molecular Weight: = ▾
[Please type Molecular Formula with atoms in **CHXNOS** (X = halogen) order]
SMILES:
Unique SMILES:
Class: All ▾
Sub-Class: All ▾
Family: All ▾

Characteristic:

□ Alkane	□ Alkene	□ Alkyne	□ Arene	□ Alcohol	□ Ether
□ Peroxide	□ Aldehyde	□ Ketone	□ Carboxilic Acid	□ Ester	□ Amine
□ Hydrazine	□ Imine	□ Nitrile/Isonitrile	□ NOx	□ Amide	□ Thiol
□ Thioether	□ Polysulphide	□ Thiocarbonyl	□ SOx	□ Halogen	□ Radical
□ Charges	□ Ionic	□ Solvation	□ Polymer		

Security code: 254671
[Type only numerical characters. Ignore letters and special characters.]

Search

Figure 4 - Screenshot of the *Structural Search* form.

3.4 System Implementation

The *ThermInfo* database was developed using MySQL and phpMyAdmin for systematic and efficient content management and administration over the web. In order to populate the database, we developed Perl scripts to parse and import the data from the spreadsheets.

The user-friendly interface, consisting of dynamic web pages, is developed using HTML, CSS, Java Script, and PHP for data visualization and management. The data is served using Apache web server and the access control is made using *.htaccess* (hypertext access).

ThermInfo is accessible from its official website: http://www.therminfo.com.

4 Results and Discussion

An overview of the *ThermInfo* database statistics in June 2009 is given in Figure 5. It contains around 3,000 unique and non redundant compounds with structural data available for all of them. The histogram examination shows us the completeness and representativeness of the dataset. We expect to see an evolution in number of structures in the next years.

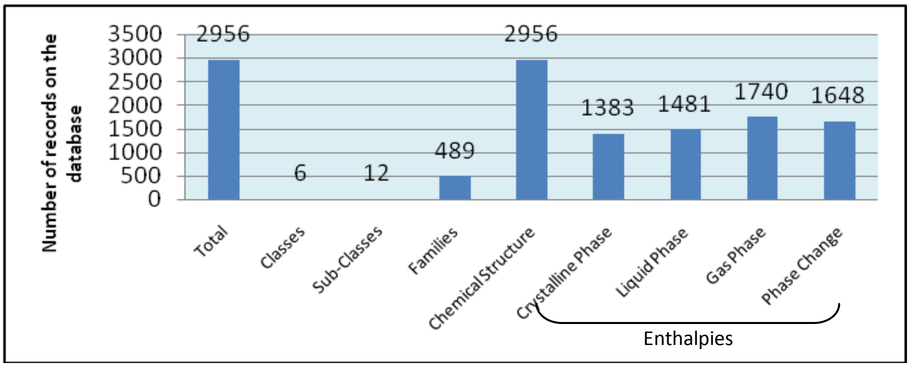


Figure 5 - Graphical representation of the database statistics with the number of records on the database in the different categories of the dataset.

Our implementation is continuously being tested by group members with different backgrounds and is under improvement based on their feedback. An initial low fidelity prototype of the *ThermInfo* interface system was developed, in order to provide directions as initial design choices for layout.

Then, with the high fidelity prototype we did a usability evaluation with 10 participants with high informatics knowledge, and 6 participants with high chemistry knowledge. To evaluate the usability of the interface, it was determined to carry out three tasks based on the features available to users of *ThermInfo*: *Simple Search* (search for a compound using a SMILES string); *Structural Search* (search for a compound based in four structural fields); and *Insert Data* (insert a new compound into de database filling five fields of the form). The evaluation focused on the time the user takes to perform the task and the number of errors committed. A questionnaire was also done, with the purpose of evaluating the easiness of the learning tasks, memorization of commands, and satisfaction in use. This questionnaire was focused in three opinion scale questions (from zero to five), each one aiming to evaluate the three points mentioned [10-11].

The results of usability evaluation are specified in Table 1. We obtained good times in performing the tasks proposed and few errors for the three tasks. The averages are similar in the two user groups. It is important to verify that for the last task (new compound insertion), although more complex than the first one (simple search), the users took shorter times to perform it and with fewer errors. This means that the similarity between the different forms/commands to realize the tasks allows a fast learning and improves the efficiency. The three subjective opinion questions, about the facility, memorization and satisfaction in use obtained very good results, all values higher than 4. All the users agreed that in future utilizations of *ThermInfo* they would not commit the same errors. All suggestions, comments and errors were taken into consideration for a new phase of interface improving.

Table 1 - Results of usability evaluation tests. The results are presented to all participants, group of information technologies experts (I) and chemistry experts (Chem).

User Group	**N**	**Task**†	**Time average (seconds)**	*St.dev.**	**Nr. of errors**	*St.dev.**	**Facility (0-5)**	**Memorization (0-5)**	**Satisfaction (0-5)**
All	16	1	66.4	18.2	0.6	0.8	4.5	4.9	4.4
		2	81.2	18.5	0.4	0.5			
		3	64.1	11.9	0.1	0.3			
I	10	1	58.9	15.6	0.3	0.5	4.3	5	4.4
		2	77.0	18.1	0.3	0.5			
		3	59.0	12.1	0.1	0.3			
Chem	6	1	77.7	16.8	1	1.1	4.8	4.8	4.3
		2	88.3	18.3	0.5	0.6			
		3	72.5	4.1	0.2	0.4			

† Task 1 is related with the *Simple Search*, task 2 with *Structural Search*, and task 3 with *Insert Data*.

* St. dev. – standard deviation of the mean value.

5 Conclusions and Future Work

This paper presented *ThermInfo*, a public web tool for collecting and presenting structural, thermochemical and bibliographic data of organic compounds. The tool was implemented in order to store the data into a relational database overcoming the problems presented by the use of spreadsheets. In addition, it provides a user and administrator interface for searching, inserting and managing the data. The development of this tool was a non-trivial task since it required collaboration between informatics and chemists.

We showed that *ThermInfo* will be a useful tool as an easy and free access source point for data fetching, which will be inevitable for chemical research and current needs. According to user surveys, it provides quick response times, as well as an intuitive and flexible interface. Compared with reputable commercial chemistry databases, *ThermInfo* is still a small database, however, its strength lies in the quality (due to the critical evaluation/validation of the data by thermochemists) and in the accessibility of the data provided.

While we will make significant efforts to screen the experimental data added, we request that the community participates in this process, expanding the thermochemical dataset. The organic compounds database compilation will be followed by organometallic, radicals, and inorganic databases. Furthermore, the integration with existing open source chemoinformatics tools, such as, JChemPaint[6] (to draw chemical structures and combine it with textual query terms to further restrict searches), and OpenBabel[7] (to convert a SMILES string in other chemical structure formats) will expand the capabilities of *ThermInfo*. As future work, we also expect the implementation and integration of prediction methods for thermochemical properties based on structure-energetics relationships [12-14].

Acknowledgments. This work was supported by FCT, through the Multiannual Funding Program. R.C.S. gratefully acknowledges Fundação para a Ciência e a Tecnologia (FCT) for a post-doctoral grant (SFRH/BPD/26610/2006).

References

1. Chen, W. L.: Chemoinformatics: Past, Present, and Future. J. Chem. Inf. Model., 46, 2230-2255 (2006).
2. Engel, T.: Basic Overview of Chemoinformatics. J. Chem. Inf. Model., 46, 2267-22774 (2006).
3. P.J. Linstrom, W.G. Mallard (Eds.), NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg MD, <http://webbook.nist.gov> (accessed in May 2009).
4. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M.: ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. Nucleic Acids Res., D344–350 (2008).
5. Weininger, D.; SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. and Comp. Sciences, 28, 31-36 (1988).
6. Weininger, D., Weininger, A., Weininger, J. L.: SMILES 2. Algorithm for Generation of Unique SMILES Notation. J. Chem. Inf. and Comp. Sciences, 29, 97-101 (1989).
7. a) Pedley, J. B., Naylor, R. D., and Kirby, S. P.: Thermochemical Data of Organic Compounds. 2nd ed., Chapman and Hall, London (1986). b) Pedley, J. B.:

[6] JChemPaint: http://apps.sourceforge.net/mediawiki/cdk/index.php?title=JChemPaint

[7] OpenBabel: http://openbabel.org/wiki/

Thermochemical Data and Structures of Organic Compounds. TRC Data Series, vol. 1, College Station, TX, (1994).

8. Modha, J., Gwinnett, A., Bruce, M.: A Review of Information Systems Development Methodology (ISDM) Selection Techniques. Omega, 18, 473-490, (1990).

9. Booch, G., Rumbaugh, J., Jacobson, I.: The Unified Modeling Language User Guide. Addison Wesley Longman, Reading, MA, 1999.

10. Dix, A., Finlay, J., Abowd, G. D., Beale, R.: Human Computer Interaction. 3rd ed., Prentice Hall, 2003.

11. Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., Carey, T.: Human Computer Interaction. Addison Wesley, 1994.

12. Poling, B. E., Prausnitz, J. M., O'Connell, J. P.: The Properties of Gases and Liquids. 5th ed., McGraw-Hill: Singapore, 2001.

13 Leal, J. P.: Additive Methods for Prediction of Thermochemical Properties. The Laidler Method Revisited. 1. Hydrocarbons. J. Phys. Chem. Ref. Data, 35, 55-76 (2006).

14. Santos, R. C., Leal, J. P., Martinho Simões, J. A.: Additivity Methods for Prediction of Thermochemical Properties. The Laidler Method Revisited. 2. Hydrocarbons Including Substituted Cyclic Compounds. J. Chem. Thermodyn., (2009), doi:10.1016/j.jct.2009.06.013.