Concept of a Robot Computational Chemist

Geerd HF Diercksen

Max-Planck-Institut für Astrophysik
Garching, GERMANY

November 27, 2007

Outline

MAX-PLANCK-GESELLSCHAFT

Outline

MAX-PLANCK-GESELLSCHAFT

Munich software

- The **Munich** software system was designed, implemented and in use from the late sixties and has served computational studies in theoretical chemistry for about 30 years.

  The software was characterised by:

  - labelled (named) data sets that can easily be located on sequensial data sets.

OpenMol software

- The **OpenMol** (*Open Mol*ecular) Program was designed, and implemented and is in use since the early ninetieth.

  The software is characterised by:

  - the use of *abstract data types* (within a Fortran environment) and
  - *modularisation* based on abstract data types (J. Guttag, 1977).

Motivation

- The ultimate goal in designing OpenMol was to build **knowledge-based** software in computational chemistry
    - to provide the **novice** (non-expert user) in academia and particularly in industry with guidance in choosing **physically correct models** which are amenable to computation
    - in order to avoid results of **unspecified reliability**.

History

- Geerd HF Diercksen
  Concepts of a Knowledge-based Simulation Environment in Physics
  in: Supercomputers, Algorithms, and Scientific Computation, FA Matsen and T. Tajima, (ed.), University of Texas Press, Austin, **1986**
- Geerd HF Diercksen and George G Hall
  Intelligent software: The OpenMol program
  Computers in Physics 8, 215 (**1994**)
- Geerd HF Diercksen
  Artificial intelligence in computational chemistry
  Research Grant Application to the
  Deutsche Forschungsgemeinschaft
  17 July **2000**

MAX-PLANCK-GESELLSCHAFT

Outline

Overall objective

- The objective of the **Robot Computational Chemist (RCC)** project is to develop and implement an artificial intelligence approach to
  - **planning**,
  - **conducting**, and
  - **understanding**

  of computational studies in quantum chemistry.

MAX-PLANCK-GESELLSCHAFT

The **Question**

- A **Question** put to the Robot Computational Chemist must contain the following *minimum* information about the subject:
    - the **quantum system** to be studied,
    - the **property** to be calculated, and
    - the **accuracy** of the property.

Outline

Overall requirements

- The requirements follow very much the steps a researcher would/**should** follow.
- They may be grouped according to the three objectives of the Robot Computational Chemist:
  - **planning**,
  - **conducting**, and
  - **understanding**

Planning

Requirements:

- checking the *domain of knowledge* of the subject,
- checking for previous studies on the subject,
- determining the necessity for a new calculation,

Planning

continued:

- selecting a suitable theoretical model,
- selecting the method for solving the model,
- selecting the software required,
- determining the computer resources required,
- checking for the availability of the required soft- and hardware resources,
- estimating the feasability of the calculation.

MAX-PLANCK-GESELLSCHAFT

Conducting

Requirements:

- compiling/editing the software to be used,
- compiling/editing the input to the software,
- submitting the calculation,
- controlling the excecution of the calculation,
- compiling the results of the calculation.

MAX-PLANCK-GESELLSCHAFT

Understanding

Requirements:

- checking the assumptions made in selecting the method and basis set,
- comparing the results to previous theoretical and experimental results.
- evaluating the reliability and accuracy of the results,
- updating the general background knowledge and case base.

Outline

Summary

- The basis for the Robot Computational Chemist is an **Evolutionary Decision Support (EDS)** system.
- The EDS architecture will be built based on a flexible methodology for knowledge-intensive **case-based reasoning**.
- Key to this approach will be the integration of general background knowledge with **specific problem-solving experience from computational chemistry**.

MAX-PLANCK-GESELLSCHAFT

Case-based reasoning

- CBR is a problem-solving model that is concerned with solving new problems by **adapting solutions that worked for similar problems in the past**. The key requirements for a case-based approach to knowledge-intensive problem-solving are:
  - a library of cases, which capture specific problem-solving experience,
  - knowledge structures that allow the identification of the most relevant cases in the light of the new problem, and
  - knowledge structures that transfer the solution from the most relevant case(s) to the new problem.

Processes

- Knowledge-intensive problem-solving involves the following processes:
  - capturing and representation of incomplete and imprecise knowledge,
  - fusion of heterogeneous information,
  - integration of contextual and general background knowledge,
  - combination of multiple-expert case bases,
  - automatic construction of case bases from data,

Processes

- continued:
  - discovery of relevant knowledge in databases,
  - discovery of case clusters for modelling and interpreting case retrieval,
  - usage of soft computing techniques for learning and construction of case retrieval knowledge structures, and
  - knowledge management.

Strategic objectives

- Three strategic objectives will serve as principal guidelines during the course of the project:
  - repository of **general background knowledge**,
  - repository of **specific problem-solving experience**, and
  - **global access and universal information system.**

EDS components

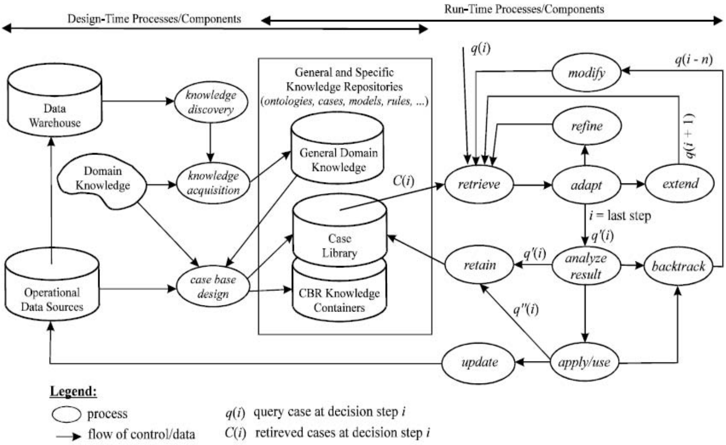

Fig. 1. Components of the EDS architecture

Outline

Repository of background knowledge

- The repository of background knowledge captures general background knowledge from **computational chemistry**.
- Such knowledge is **independent** from any specialised problem-solving or decision-making knowledge.
- This approach allows to use the knowledge repository in a variety of ways.

Repository of background knowledge

- The repository of background knowledge provides the basis for
  - efficient and effective mechanisms for retrieval of relevant information such as knowledge artefacts (e.g. cases and rules) in a knowledge-based system or documents on the Web,
  - systems used in **education** and **training**, and
  - a knowledge navigation system that **experts** and **novices** can use directly to **explore concepts in chemistry**.

Repository of specific problem-solving experience

- The repository of specific problem-solving experience captures specific problem solving **episodes** or **cases** and provides suitable query and retrieval mechanisms.
- Using analogy-based access methods and knowledge-based adaptation techniques, the cases in this repository may be **re-used** in a wide variety decision-making scenarios.
- Crucial for the effective re-use of this knowledge source will be its **integration** with the more general knowledge structures of the repository of background knowledge.

Global information system

- Initially, the global access and universal information system will be implemented as a **localised, isolated** decision support facility.
- In future, it is intended to extend the systems basic architecture so that its knowledge sources (general and specific) can be embedded as so-called infohabitants in a future **universal** information system.
- This vision includes passive access (decision support) as well as largely autonomous forms of pattern formation and self-organisation such as concept learning.

EDS components

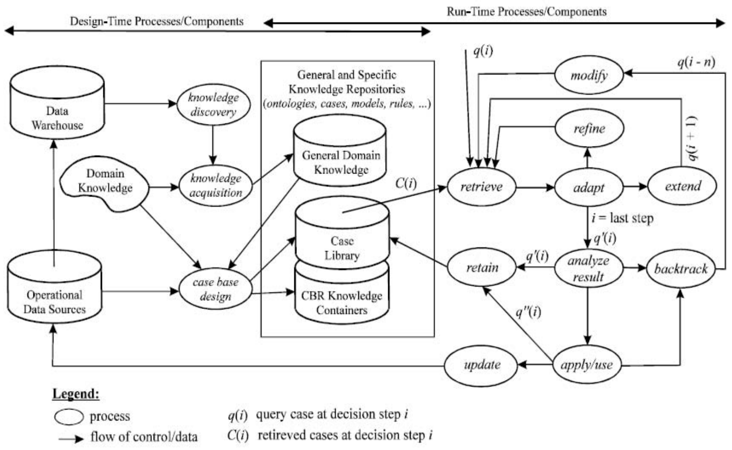

Fig. 1. Components of the EDS architecture

Outline

MAX-PLANCK-GESELLSCHAFT

Objective

- The EDS framework describes the processes involved in problem-solving through a highly modular architecture of knowledge structures and repositories, such as
  - general,
  - specific,
  - retrieval,
  - adaptation knowledge, ...

Objective

- The core processes reflected in the EDS architecture include:
  - **retrieval** of relevant knowledge artefacts such as problem-solution episodes, ontology partitions, rules, etc.,
  - transfer or **adaptation** of the information contained in these artefacts in the light of the new problem,
  - **evaluation** of the proposed course of action or solution, and
  - **integration** of the new problem-solving episode into the knowledge repositories.
- A key concept in this architecture is that these processes may take place in an **iterative** fashion permitting the evolution of a solution.

MAX-PLANCK-GESELLSCHAFT

EDS components

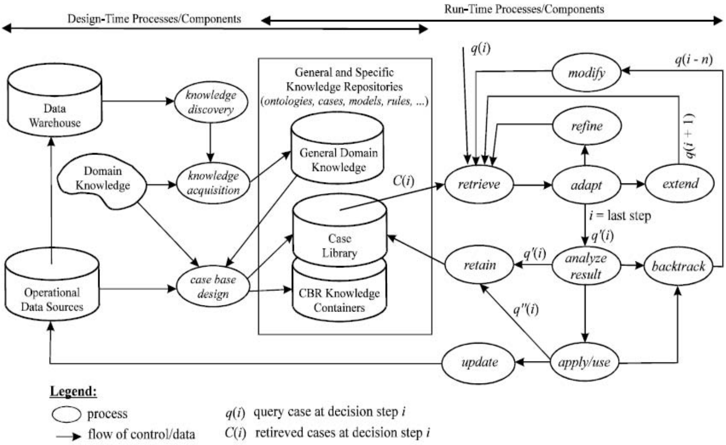

Fig. 1. Components of the EDS architecture

General

The implementation is organized into:

- 3 tracks and
- 6 workpages.

MAX-PLANCK-GESELLSCHAFT

General

Critical to the project will be a

- **structured** approach to information and knowledge modelling with
- a clear concept for **modularity** and **openness** (universal information system).

Therefore,

- **object-oriented** tools and techniques and
- the **CommonKADS** methodology

will be used as key methodical backbone for the project to ensure a systematic development.

General

The 3 tracks are:

- acquisition, capturing, and encoding of **actual** knowledge structures (general and specific knowledge components),
- development of the **generic** knowledge repositories and access protocols,
- development of the **generic** EDS engine.

Because of the **decoupling** of knowledge repositories and decision processes, some of the development will initially proceed in parallel along three main tracks.

General

The 6 work packages address the following topics:

- problem **characterisation** and initial knowledge **acquisition**,
- knowledge **acquisition** and engineering: general knowledge,
- knowledge **repositories** and access,
- evolutionary **decision support engine**,
- **integration** of knowledge into repositories of the overall system,
- **integration** of numerical software.

MAX-PLANCK-GESELLSCHAFT

General

The work package *Knowledge acquisition and engineering: general knowledge* breaks down into 4 sub-packages:

- knowledge repository: background knowledge,
- knowledge repository: case library,
- knowledge repository: retrieval knowledge,
- knowledge repository: adaptation knowledge structures.

Problem characterisation and initial knowledge acquisition

- This work package is concerned with **identifying and describing the key knowledge and information areas, structures and resources** (cases, concept and problem spaces, ontologies, databases, etc.) needed to develop a robust and competent system.
- This phase will see a **very close and intensive collaboration** between the **domain experts** and the **knowledge engineers**.

MAX-PLANCK-GESELLSCHAFT

Knowledge acquisition and engineering: general knowledge

- This work package is concerned with **eliciting, capturing, and encoding** of the knowledge needed to populate the various knowledge repositories.
- Based on the initial knowledge acquisition effort, key activity in this work package will be concentrated on the **acquisition** of **detailed general and specific knowledge structures** used within the chosen subset of **computational chemistry**.

MAX-PLANCK-GESELLSCHAFT

Knowledge repositories and access

- This work package is concerned with **designing, implementing and populating the actual knowledge repositories and access protocols** using adequate design and implementation methodologies, platforms, and languages.
- The methodology used in this work package will be strongly based on **object-oriented analysis, design and implementation principles**, tools and languages.

Evolutionary decision support engine

- This work package is concerned with **designing and implementing the generic EDS engine** using adequate design and implementation methodologies, platforms, and languages.
- Key tasks within this package are concerned with the development of algorithms that support the main processes of the EDS system:
  - retrieval,
  - adaptation,
  - evaluation, and
  - integration.

Integration of knowledge into repositories of the overall system

- This work package is concerned with **integrating the acquired knowledge** into the corresponding repositories of the EDS engine, and test and evaluate the resulting system.

Integration of numerical software

- This work package is concerned with **integrating existing numerical software** used in computational chemistry into the prototype system, in particular the OpenMol software specifically developed for this purpose.

Knowledge acquisition - Knowledge repository: background knowledge.

- This sub-package is concerned with *identifying, capturing, characterising, and encoding* an **adequate subset of general knowledge types and concepts** (ontologies, guidelines, computation parameters, rules, classes, etc.) from chemistry that will form the basis for the system.

MAX-PLANCK-GESELLSCHAFT

Knowledge acquisition - Knowledge repository: case library

- This sub-package is concerned with *identifying (generate if necessary), capturing, characterising, and encoding* an **adequate set of concrete problem-solving episodes or cases**, which can be understood (described) in sufficient detail by the body of available general knowledge.
- This requires the development of a suitable case representation format that
  - captures the knowledge necessary to derive answers put to the system, and
  - is amenable to the general EDS process model.

MAX-PLANCK-GESELLSCHAFT

Knowledge acquisition - Knowledge repository: retrieval knowledge

- This sub-package is concerned with *identifying, capturing, and encoding* **adequate knowledge structures** that will reliably and efficiently identify full or partial cases within the case library given a set of descriptors.

Knowledge acquisition - Knowledge repository: adaptation knowledge structures

- This sub-package is concerned with *identifying, capturing, and encoding* **adequate knowledge structures** that will reliably and efficiently **analyse and modify** (adapt) the retrieval results at intermediate steps of the evolutionary decision process.
- These structures will also hold the knowledge required for **evaluating the intermediate results** (decisions, solutions) obtained in the various decision processes.

Outline

Similarity

- The open question in this project that needs some basic research efforts concerns **similarity**:
- *What really defines similarity of quantum systems with respect to the methodology suitable for studying them?*

Summary

- The development and implemention of a **local pilot version of restricted scope** of the EDS was layed out as
  - a **2 year project**
  - for **2 post-doctoral fellows**.
- The refrees of the DFG argued that time and manpower is underestimated considering the objectives of the project. Probably they are right.

MAX-PLANCK-GESELLSCHAFT

Outline

Individuals

Major contributions to this project over the past two decades by the following colleagues are greatefully acknowledged:

Werner Dubitzki
Wlodek Duch
Jacek Karwowski
Al Matsen
Stan Scott
Mike Zerner

Acknowledgement: MPI für Astrophysik

MAX-PLANCK-GESELLSCHAFT

MAX-PLANCK-GESELLSCHAFT