ity mapping that finally confirms (or not) the presence of the required query substructure.

24.3.2.7. *Data validation*

All data entering the CSD are subject to stringent check and evaluation procedures. Some of these are visual, but the majority are automated within the CSD program *PreQuest*. The checks ensure that the 1D and 2D information fields abstracted by CCDC staff are accurately encoded, and that the 3D crystallographic coordinates are consistent with both the chemical description of the structure and with the geometrical description supplied by the authors. Most typographical errors in original papers can be corrected by the CCDC but, in the case of serious discrepancies, the original authors are consulted.

24.3.2.8. *The CSD-Use database*

CSD-Use is a database of scientific research papers in which the CSD was used as the principal or sole source of experimental information. The database comprises more than 700 literature citations classified according to the type of systematic study undertaken. Each CSD-Use entry also contains a short summary of the major findings of the research. The database is growing rapidly over time, and is expected to be a valuable resource in the future, since it contains a fully retrospective overview of the data-mining methods and research applications of the CSD.

24.3.3. The CSD software system

24.3.3.1. *Overview*

The CSD is supplied with a suite of fully interactive graphical software modules which provides users with facilities to: (*a*) interrogate all of the 1D, 2D and 3D information fields; (*b*) display entries graphically in a variety of styles; (*c*) retrieve relevant data for search hits, including geometrical parameters derived from the stored coordinates; and (*d*) display the derived numerical information, *e.g.* as histograms, scattergrams *etc.*, generate descriptive statistics and perform more complex numerical analyses. More recently, software has been added that permits users to transform their own in-house structural data to CSD formats for inclusion in these processes. A summary of the overall CSD software system is given in Fig. 24.3.3.1 which shows the functional relationships between the four major applications programs.

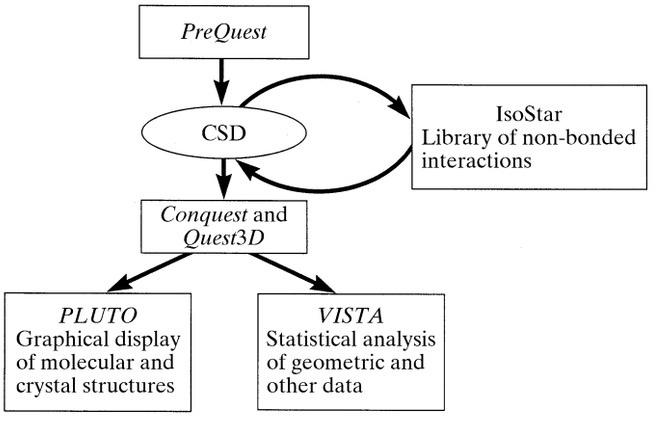

Fig. 24.3.3.1. Summary of the software components of the Cambridge Structural Database system (CSDS).

24.3.3.2. *PreQuest*

PreQuest is a data-validation and data-conversion program which is used to create high-quality structural data files in CSD format from, *e.g.*, raw input data from a CIF. *PreQuest* is used routinely by CCDC's scientific editors to create and validate entries for inclusion in the master CSD archive, hence the program is constantly being maintained and upgraded. The released version enables users to build a private CSD-format database of their own structures which can then be searched independently of, or in conjunction with, the master CSD files using the database access programs described below.

24.3.3.3. *Searching the CSD: Quest3D and ConQuest*

Quest3D has been the main search engine and information-retrieval program for the CSD since the late 1980s. Its main features are summarized below. However, since 1997, the CCDC has been developing its successor, the *ConQuest* program, which was first released as part of the CSD system in April 2000. During an interim period, perhaps two years, *ConQuest* and *Quest3D* will both form part of the released CSD system on certain computing platforms while the functionality of the new program is being fully developed. Further details of *ConQuest* are provided in Section 24.3.3.5, indicating in particular how it differs from, and improves upon, the facilities available in *Quest3D*.

24.3.3.4. *Quest3D*

Quest3D is the main search engine and information-retrieval program for the CSD. It permits interrogation of all information fields: (*a*) 19 text fields, (*b*) 38 individual numerical fields, (*c*) element symbols and element counts, (*d*) full or partial molecular formulae, (*e*) direct access to over 150 bit screens, (*f*) extensive 2D chemical substructure search capabilities, and (*g*) 3D substructure searching at the molecular level or at the extended crystal-structure level. A search of a specific information field is termed a *test* of that field, and is constructed graphically *via* the menu system; menu components correspond to the categories of searches identified above. A complete *query* is then constructed by combining a number of separate *test* components using Boolean logic.

Substructure searching is the most important and frequently used facility. At the molecular level, the substructure (chemical fragment) query is entered graphically and is defined using the formal covalent bond types present in the 2D chemical connectivity tables of the CSD. The process can be extended to locate non-bonded contacts in the complete crystal structure. Here, the individual atoms or chemical groups involved in the contact must be specified, and a limiting non-bonded contact distance must be provided, along with any other geometrical criteria required to define the contact more precisely.

All substructure searches begin with the user drawing the required chemical unit(s) *via* the BUILD menu. Chemical variability and precision are controlled through (*a*) the PERIODIC TABLE sub-menu, which allows for specification of variable element types at specific atomic sites, (*b*) the 2D-CONSTRAIN menu, which allows further chemical restrictions to be specified, such as cyclicity/acyclicity of bonds, exact hydrogen-atom counts, total coordination numbers for atoms *etc.*, and (*c*) the 3D-CONSTRAIN menu, which permits the user to specify a list of geometrical parameters to be calculated by the program for each instance of the fragment located in the CSD; any of these geometrical parameters may be used as criteria to limit the scope of the search, especially at the intermolecular level. A file of calculated geometrical information is output by *Quest3D* and may be read by *Vista*, or by external data analysis software. Other

references