equivalent observed intensities around the average value. However, the most popular form of $R_{\rm merge}$ given above is not a proper, statistically valid quantifier. It does not take into account the multiplicity of the measurements and, as a consequence, it actually rises with increased multiplicity, falsely indicating degradation of the data quality when in reality they have a higher accuracy. Modifications of $R_{\rm merge}$ have been proposed to include the effect of multiple measurements properly (Diederichs & Karplus, 1997; Weiss, 2001) (see Chapter 2.2).

A better quantity for assessing the quality of the X-ray data is the $\sum_{hkl} I_{hkl}/\sum_{hkl}\sigma(I_{hkl})$ ratio, provided the standard uncertainties, $\sigma(I)$, are correctly estimated. Detectors such as imaging plates or CCDs do not measure individual X-ray quanta directly, having a gain factor dependent on the response of the individual detector pixel to a single X-ray photon. If the gain factor is not known accurately for a particular detector, the resulting standard uncertainties of the measured intensities will be estimated at an incorrect level. If the multiplicity of the reflections is higher than unity, it is possible to correct the uncertainties *a posteriori*. This can be done either from a comparison with the expected values using the χ^2 test, or by using the *t*-plot. The latter requires that the ratio of the differences between equivalent intensity measurements to their standard uncertainties, $t = (I_i - \langle I\rangle)/\sigma(I_i)$, follows a normal distribution with a mean of 0.0 and standard deviation of 1.0. Both of these methods assume the errors have a normal distribution, and that only the mean and width have been incorrectly estimated and should be appropriately adjusted. They cannot take into account systematic errors of measurement.

The data-merging procedure in addition allows the identification of statistical 'outliers' and their exclusion from the data (Read, 1999). Outliers are defined as those observations that lie sufficiently far from the mean of a set, and assumption of a normal distribution suggests they suffer from substantial systematic errors of measurement. In a crystallographic experiment, outliers are those intensity measurements that deviate unexpectedly from the mean intensity of a set of symmetry-equivalent reflections. In the recording of rotation data, one typical source of such systematic errors is erroneous classification of reflections predicted as partially or fully recorded. This is a severe problem for those reflections lying close to the blind region. A second example is the presence of so-called 'zingers' in individual CCD detector pixels caused by scintillations from trace radioactivity of the taper glass. Other problems such as shadowed or inactive regions of the detector window give rise to a range of such systematic errors.

A small number of outliers may be expected from such causes. However, the total fraction of reflections flagged as outliers and rejected from the merging process should be small, certainly much less than 1%. Larger fractions indicate serious deficiencies in the hardware or the software and suggest something is very wrong with the experiment. There should always be a physical reason for rejecting outliers, other than just a need to reject those agreeing poorly with their symmetry-equivalent intensities in order to drive down $R_{\rm merge}$. It is always possible to reduce $R_{\rm merge}$ and to provide an apparent 'improvement' in the data by rejecting a large percentage of measurements, but this is extremely bad practice.

Good crystallographic data depend strongly on an appropriate statistical procedure. It is also inappropriate to exclude those reflections with intensities lower than a cutoff limit, such as 1σ, before or during the process of data merging. Weak intensities also carry information and their neglect introduces bias into the measured intensity distribution, affecting, for example, the overall or individual atomic temperature factors.

The true outer resolution limit of the diffraction pattern is not trivial to define and indeed depends to some extent on the application. If $I/\sigma(I)$ is higher than 1.0, then a resolution shell of data indeed contains some information in a statistical sense – provided of course that $\sigma(I)$ has been correctly estimated. However, as $I/\sigma(I)$ falls close to unity there will in practice be very few significant observations amongst a great deal of noise. It is necessary to make some decision about where to cut the effective resolution. For the application of direct methods, for example using *SHELXD* (Sheldrick, 2008), the cutoff is often defined as the resolution shell where $I/\sigma(I)$ falls to 2.0, when $R_{\rm merge}$ usually reaches 20–40% depending on the symmetry and redundancy. Cruickshank (1999*a*,*b*) has provided a formula for a data precision indicator (DPI) which includes the effect of falling $I/\sigma(I)$ ratio (see Chapter 2.2).

For other applications it may be advisable to accept even very weak data. Direct methods use only a subset of the most meaningful reflections, but which should extend to as high a resolution as possible. In addition, when the data are sparse from crystals that only diffract to very limited resolution, perhaps around 3 Å, then it is essential to retain all the experimental data, even if they are weak.

9.1.12. Radiation damage

9.1.12.1. Historical perspective

All crystals irradiated with X-rays absorb at least a fraction of the radiation, resulting in damage to the sample (Henderson, 1990). The energy from the absorbed photons may initially result in the disruption of chemical bonds, before being eventually dissipated as thermal energy. For well ordered small-molecule crystals the lattice is close-packed and the effects arising from the absorbed photons are restricted to the immediate environment of the absorption event, so-called primary damage. Only when a substantial fraction of the crystal has been affected do cooperative effects set in.

In contrast, roughly 50% of a macromolecular crystal is disordered aqueous solvent (Matthews, 1968). At room temperature this allows a secondary mechanism of radiation damage, resulting from diffusion of radicals and ions produced at the primary absorption site which affect chemical moieties at positions remote from this site. The details of this process remain poorly understood but are related to the extremely damaging effects of X-rays on biological tissue. A consequence of this damage is that degradation of the crystal order continues even after the irradiation is stopped or interrupted. For collection of data at room temperature from protein crystals mounted in capillaries, secondary damage contributes significantly to the rate of deterioration of the diffraction pattern. One of the gains of the early applications of SR was that it allowed recording of data to proceed ahead of the effects of secondary damage, increasing the effective, if not the absolute, lifetime of the crystal in the X-ray beam. An experiment often required several crystals, all of which showed the effects of temporal decay in their recorded intensities, which needed to be merged to provide complete data.

9.1.12.2. Cryogenic vitrification

In the early 1990s, the introduction of protein-data collection at cryogenic temperatures, using so-called flash cooling, was a major breakthrough (Garman & Schneider, 1997; Rodgers, 1997;

Garman & Owen, 2006). Such vitrification of crystals largely prevented the effects of secondary damage. On the X-ray sources then available, it was in most cases possible to record complete data from a single sample without significant degradation of the diffraction, enormously simplifying the strategy of data collection and merging.

Almost all data are currently collected from vitrified samples (see Part 10). The prolonged life of the sample and modest rates of data acquisition, even at second-generation SR sources with imaging plates, allow enough time for careful analysis of the initial images and optimization of the strategy.

A second major advantage of cryogenic data collection is that it allows crystals to be reused after initial data have been recorded. Two examples show the usefulness of this approach. Firstly, when screening the binding of heavy atoms for phase determination or ligands for complex formation, data can first be recorded to the minimum resolution needed to determine whether the binding is successful. Secondly, a series of vitrified crystals can be screened for their degree of order in the home laboratory, and the best stored and retained for subsequent improved collection either in the home laboratory or at a synchrotron site. The ability to transport vitrified crystals has proved invaluable in this respect, and leads to optimal use of synchrotron resources.

9.1.12.3. High-intensity third-generation SR sources

The advent of third-generation SR sources and insertion devices has led to X-ray beams of unprecedented intensity. The speed of data collection can be of the order of 1 second per 1° rotation. In association with CCD detectors able to read out images within a few seconds, this means that a complete data set can be obtained in a few minutes. At first sight, this would seem to have solved the problem of macromolecular data collection, as such speeds should allow recording of highly redundant accurate data to the highest resolution in a tractable time.

However, with such high intensities it appears that the effects of radiation damage are significant and result in specific effects on susceptible parts of the structure. The useful active exposure lifetime of typical crystals seems to be around five minutes, with substantial degradation of the diffraction pattern ensuing even for vitrified crystals. The first manifestation of radiation damage is the disruption of disulfide bridges and decarboxylation of aspartates and glutamates. This effect means appropriate strategies for selecting the optimal starting point of rotation in order to minimize the total rotation required for collection of complete data are once more essential. Several strategy programs, such as *BEST* (Popov & Bourenkov, 2003; Bourenkov & Popov, 2006), now permit this to be done effectively.

9.1.12.4. Correcting data for the effects of radiation damage

The overall effect of radiation damage is that the higher-resolution intensities decrease faster than those at low resolution. This effect is largely taken into account by the relative *B* factors applied to individual images during data scaling and merging by the major data-reduction programs.

However, such scaling does not allow for the effect of specific structural damage (*e.g.* the S–S bridges and carboxylic groups) on individual reflection intensities. A method to deal with this has been proposed by Diederichs *et al.* (2003). This is based on a zero-dose extrapolation of intensities and requires that a time-stamp be attached to each individual intensity measurement. Such a timestamp has also been used to assist in estimation of phases by *SHARP* (Schiltz *et al.*, 2004).

9.1.13. Relating data collection to the problem in hand

The data-collection protocol should be matched to the purposes for which the data are to be used. Different applications present a range of different needs, requiring the intensities (or structure-factor amplitudes) to be exploited in different ways. In this section a representative set of applications is outlined in terms of how the tactics and strategies of data collection can vary.

9.1.13.1. Isomorphous-anomalous derivatives

The phasing of proteins by isomorphous replacement requires the collection of data from crystals of one or more heavy-atom derivatives of the protein that are isomorphous to the parent native crystal. Preparation of derivatives involves either soaking of native crystals in the heavy-atom solution or co-crystallization with the heavy-atom reagent (Part 12). Data collection can be split into two parts. The first step is to establish whether a potential derivative is isomorphous and contains the expected heavy atoms. The second is to collect the data on this derivative to provide the necessary phase information for the native structure factors. The problems of how to utilize the phase information are addressed in Part 12. Here, strategies applicable to the two steps are described.

Screening of derivatives can be carried out by collecting data to the resolution limits of the crystals. This can consume substantial data-collection resources and lead to irrelevant data that are not from isomorphous crystals or do not contain the anticipated heavy-atom signal. It is preferable to record the minimum data sufficient to identify a potential derivative in order to save time and resources, as many samples may need to be screened. A minimal strategy can exploit some or all of the following protocols:

(1) An essentially complete native-data reference set should be available, although not necessarily to the ultimate resolution limit.
(2) Preparation of a set of crystals with a selected set of potential heavy atoms, the number depending on crystal availability.
(3) Collection of a small number of images from each potential derivative crystal, ideally on the home-laboratory rotating-anode source or an SR beamline if necessary. These data can be recorded to a low resolution: in principle 4 Å or less should be enough. The resulting partial derivative data are scaled with the complete native set. The fractional isomorphous difference can be evaluated easily and compared with the expected agreement with the native data. In general, values less than 10% suggest that the heavy atom is not bound. Values higher than about 30% suggest an unacceptable level of non-isomorphism. Intermediate values suggest, but do not guarantee, that the derivative is worth pursuing. Normal probability plots can be helpful in this respect (Howell & Smith, 1992).
(4) Given a positive result from point (3), complete data may be recorded on the same or an equivalent derivative crystal. Again, it may be useful to record data to low resolution in the first instance. 4 Å resolution is again quite sufficient to solve the structure of a heavy-atom constellation using direct or Patterson methods, allowing the more complete characterization of the potential derivative.