Science from Gaia: how to deal with a complex billion-source catalogue and data archive

Anthony G.A. Brown

Abstract The Gaia mission will provide us with an unprecedented stereoscopic map of the heavens and will likely be *the* astronomical data resource for decades thereafter, representing a tremendous discovery potential. I will summarize the Gaia mission and the expected catalogue contents and then show how the complexities of the catalogue, and the science we want to extract from it, will force us to be very ambitious in the way we publish the Gaia catalogue. Truly unlocking its potential requires integrating the Gaia catalogue with other sky surveys and using advanced statistical approaches to extracting the science, ultimately aiming at facilitating hypothesis testing against the raw image pixels collected by Gaia.

1 The Gaia mission

Gaia is the European Space Agency mission which will carry out an all-sky astrometric, photometric, and spectroscopic survey — observing every object brighter than 20^{th} magnitude — amounting to about 1 billion stars, galaxies, quasars and solar system objects. Gaia is scheduled for launch in 2013 and over the course of its five year survey will measure positions, parallaxes, and proper motions with expected accuracies of 10–25 μas, depending on colour, at 15^{th} magnitude and 100–300 μas at 20^{th} magnitude. The astrometric measurements are collected employing a wide photometric band (the Gaia G band) which covers the range 330–1000 nm. Multi-colour photometry will be obtained for all objects by means of low-resolution spectrophotometry. The photometric instrument consists of two prisms dispersing all the light entering the field of view. One disperser — called BP for Blue Photometer — operates in the wavelength range 330–680 nm; the other — called RP for Red Photometer — covers the wavelength range 640–1000 nm. In addition radial veloc-

Anthony G.A. Brown
Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA Leiden, The Netherlands, e-mail: brown@strw.leidenuniv.nl

ities with a precision of 1–15 km s^{-1} will be measured for all objects to 17$^{\rm th}$ magnitude, thus complementing the astrometry to provide full six-dimensional phase space information for the brighter sources. The radial velocity instrument (RVS) is a near-infrared (847–874 nm, $\lambda/\Delta\lambda \sim 11\,000$) integral-field spectrograph dispersing all the light entering the field of view. Gaia builds on the proven principles of the Hipparcos mission [1] but represents an improvement of several orders of magnitude in terms of numbers of objects, accuracy and limiting magnitude (Hipparcos observed 120 000 stars to 12$^{\rm th}$ magnitude, achieving milli-arcsecond accuracy).

The scientific power of Gaia rests on the combination of three desirable qualities in a single mission: (1) the ability to make very accurate (global and absolute) astrometric measurements; (2) the capability to survey large and complete (magnitude limited) samples of objects; and (3) the matching collection of synoptic and multi-epoch spectrophotometric and radial velocity measurements (cf. [2]). The range of science questions that can be addressed with such a data set is immense and Gaia will surely revolutionize almost every field in astronomy, including the study of the very distant quasars and the very nearby solar system asteroids. I will not attempt to summarize the Gaia science case here but point to the proceedings of the conferences *The Three-Dimensional Universe With Gaia* [3] and *Gaia: At the Frontiers of Astrometry* [4]. More detailed information on the scientific performance numbers for Gaia can be found on-line at `http://www.rssd.esa.int/index.php?project=GAIA&page=Science_Performance`.

1.1 Gaia catalogue contents

In order to set the stage for the discussion in the rest of this paper it is interesting to consider what primary scientific information the final Gaia catalogue will contain. In Gaia's own broad-band magnitude G the number of stars in the catalogue is estimated to be $\sim 7\times10^5$ to $G=10$, 48×10^6 to $G=15$ and 1.1×10^9 to $G=20$. About 60 million stars are expected to be seen as binary or multiple systems by Gaia, among which about 10^{6-7} eclipsing binaries. For each source observed by Gaia the following information is provided:

astrometry positions, parallax, proper motions, the full covariance matrix of the astrometric parameters (standard errors and correlations) and astrometric solution quality indicators.

photometry broad band fluxes in the G, $G_{\rm BP}$, $G_{\rm RP}$ and $G_{\rm RVS}$ bands, as well as the prism spectra measured by the blue and red photometers. Variability indicators will be provided for all stars together with epoch photometry.

spectroscopy radial velocities for the $\sim 150\times10^6$ stars at $V\leq17$; rotational velocities ($v\sin i$), atmospheric parameters, and interstellar reddening for the $\sim 5\times10^6$ stars at $V\leq13$; abundances for the $\sim 2\times10^6$ stars at $V\leq12$; accumulated spectra for the stars at $V\leq13$. The spectroscopic data are expected to contain about 10^6 spectroscopic binaries and about 10^5 eclipsing binaries.

multiple stars solution classifications and, where relevant, orbital parameters together with covariance matrices and quality flags.

astrophysical parameters the Gaia catalogue will provide as much astrophysical information on each star as possible, derived from the combination of photometric, spectroscopic and parallax information. The astrophysical parameters include $T_{\rm eff}$, A_V, $\log g$, [M/H], and [α/Fe] where possible. Luminosities and ages will also be provided (see contributions by Smith, Liu and Tsalmantza in this volume).

variability survey for about 10^8 stars a variability analysis will be provided and estimates indicate that about 20×10^6 classical variables and $1–5 \times 10^6$ eclipsing binaries will be found, among which ~ 5000 Cepheids and 70 000 RR Lyrae.

In addition the catalogue will contain astrometry and photometry for $\sim 3 \times 10^5$ solar systems bodies, $\sim 5 \times 10^5$ quasars, and some 10^6–10^7 galaxies.

This would clearly be an overwhelming data set to deal with if it were to land on one's desk today. We should thus definitely prepare carefully if we want to make full use of the Gaia catalogue data. In the following I will review the pitfalls of working with survey data such as provided by Gaia and discuss, using the example of modelling the Milky Way galaxy, the complications we will face when attempting to use the Gaia catalogue to answer a science question. This will lead to a number of proposals regarding the publication of Gaia results that are aimed at ensuring that we can make optimal use of the Gaia survey well into the future, including the combination with other existing and future large sky surveys. In this discussion I also try to identify the research that needs to be done in order to guarantee the optimal scientific exploitation and future preservation of the Gaia catalogue and data archive.

2 Extracting science from Gaia; pitfalls and complications

I discuss here a number of the pitfalls that should be taken into account when dealing with a very large survey such as provided by Gaia. The emphasis will be on the problems in dealing with an astrometric survey but many of the issues are generic to surveys in general. To illustrate the complications of extracting optimal science from the Gaia data I will discuss the example of building a self-consistent model of our Galaxy that is capable of 'explaining' the Gaia catalogue data.

2.1 Effects complicating the interpretation of the Gaia survey

The main effects complicating the interpretation of the Gaia survey data are:

Completeness and selection effects Although the Gaia survey is designed to be complete and unbiased to $G = 20$, the details of the on-board detection software, the survey strategy (the 'scanning law', see contribution by Holl in this volume),

and data loss due to mission interruptions and loss of telemetry packets, will lead to varying detection and completeness limits over the sky. In particular In high density regions ($\geq 10^5$ sources per degree^2) the effective magnitude limit may be brighter or the number of observations per source smaller than the average. The statistical description and analysis of the varying completeness of the Gaia catalogue will be a delicate issue to deal with.

Correlated errors In general the errors of the astrometric parameters for a given source will not be statistically independent. Moreover the errors for *different* sources may also be correlated. The latter case is described in more detail in the contribution by Holl in this volume. The covariance of the errors for a given source will be provided in the Gaia catalogue and should be used. Ignoring these correlations may lead to spurious features in the distribution of derived astrophysical quantities. For examples of such features in the Hipparcos data see [5].

Systematics as a function of sky position The details of the way in which the Gaia measurements are collected (revolving scanning of the sky along great circles, using two telescopes) will be reflected in systematic variations of the errors and their correlations over the sky. Taking these systematics into account is especially important for studies that make use of sources spread over large sky areas.

Estimating astrophysical quantities When estimating astrophysical quantities from the analysis of samples of objects it is natural to first calculate these quantities for each individual object from the astrometric (and complementary) data and then analyse their distribution in the space of the astrophysical parameters. This allows us to work with familiar quantities such as distance, velocity, luminosity, angular momentum, etc. However, it is important to keep in mind that the actual data do not represent the astrophysical parameters in their natural coordinates. In particular it is *not* the distance to sources which is measured directly but their parallactic displacements on the sky caused by the motion of the earth around the sun (listed as the parallax ϖ in the catalogue). As a consequence many astrophysical quantities are non-linear functions of the astrometric parameters. Examples are the distance itself and the absolute magnitude which are functions of $1/\varpi$ and $\log \varpi$, respectively. Simplistic estimates of astrophysical parameters from the astrometric data can then lead to erroneous results. The only robust way around this is forward modelling of the observables or the data, as discussed in section 3.1

3 A model of our Galaxy to explain the Gaia catalogue

The main science driver for Gaia is the unravelling of the structure and formation history of the Milky Way. The Gaia catalogue can of course be used to carry out straightforward studies of specific Galactic components (thin and thick disks, bulge, bar, halo) in order to characterize them to high precision. However with the opportunities provided by Gaia we should be much more ambitious. The structural components of our Galaxy are coupled through gravity and the observed stellar and

gas kinematics are determined by the gravitational potential of the Galaxy. The only way to develop a consistent understanding of the mass distribution and kinematics is through a dynamical model of the Galaxy, and it is only with such a model that one can make reliable extrapolations to the unobserved parts of Galactic phase space. The Gaia catalogue can be seen as a snapshot of the state of the Galaxy in which we will be seeing stars from the same population at different points along the same orbits. This allows the reconstruction of individual orbits from which we can infer the Galactic potential and matter distribution. Any dynamical model will thus be highly constrained.

In addition the model of our Galaxy should also be able to explain the stellar populations in the Galaxy and thus make predictions for their distributions in age, luminosity, metallicity, and chemical abundance patterns. Hence as argued in [6], if we want to take full advantage of an all-sky high accuracy astrometric data-set, complemented by radial velocities, photometry and astrophysical information, and convert this data for 1 billion stars into a complete physical understanding of the structure of our galaxy, the goal should really be to construct a model in terms of which we can explain the data contained entire Gaia catalogue.

Constructing such a model is obviously a non-trivial task. The model has to be able to self-consistently determine matter and velocity distributions from the underlying potential. Moreover, in comparing with the Gaia catalogue data the astrophysical properties of the stellar populations have to be explained as well and the effects of extinction due to dust have to be accounted for. Several options for preparing such models are discussed in [6],[7] and [8].

3.1 Finding the best Galaxy model

Whatever modelling approach one chooses, one is faced with the enormous task of deciding which Galaxy model is best through a comparison to the rich data contained in a billion star catalogue. The basic predictions from the models are the distributions, at some time, of the stars in phase-space $(\mathbf{r},\mathbf{v})$ and in the space of astrophysical parameters (magnitude, colour, $\log g$, [M/H], [α/Fe], age, ...). The natural approach would be to take the observational data contained in the catalogue and convert those into the data-space of the model. This approach suffers from several problems:

- The effects of dust in the interstellar medium have to be corrected for.
- For most stars the radial velocity will not be available which will lead to incomplete phase space information. The interpretation is not trivial as only velocities perpendicular to our line of sight are then known. An example of how to deal with a lack of radial velocities when interpreting the velocity distribution of stars in the solar neighbourhood can be found in [9] and [10].
- As mentioned in section 2.1 the simplistic estimation of luminosities, distances, and transverse velocities from the observed photometry and astrometry can lead to erroneous results. For example the familiar integrals of motion, energy E and

angular momentum L, are functions of $1/\varpi^2$. The energy-angular momentum plane is a powerful tool when looking for remnants of accreted satellites. However as shown in [11] the propagation of the parallax errors can lead to sign changes in L_z and spurious caustic structures in the integrals of motion space that may be mistaken for physical entities

- As mentioned in section 2.1 the errors on the various quantities in the Gaia catalogue will vary over the sky and are correlated, including correlations from star to star. The non-linear relation between parallax and quantities derived from it will, when converting the observations to the model space, lead to strongly non-Gaussian errors with complicated correlations between them.

Hence, as recently also argued in [8] the complications introduced when converting the observations into intuitively more easily understood quantities will make it almost impossible to achieve a satisfactory understanding of how observational errors relate to the uncertainties in our model parameters. As a consequence deciding on the 'best' model will become impossible.

The only truly robust way to get around this problem is to project the Galaxy model into the data-space (i.e., use 'forward modelling') and thus predict the astrometric data together with the other data in the Gaia catalogue (radial velocities, magnitudes, colours, and astrophysical parameters the of stars). The added advantage is that one can readily account for incomplete phase space data (e.g., lack of radial velocity data) and selection effects. The extinction due to dust can be taken into account in predicting the observed distribution of magnitudes and colours of the stars. Moreover, negative parallaxes (which are perfectly legitimate measurements!) and the correlations in the errors on the astrometric parameters (which will vary systematically over the sky) can be much more easily accounted for in the data-space. Finally, the ongoing discussions in the literature on the Lutz-Kelker 'bias' and how to deal with it (e.g., [12], [5]) can be entirely avoided by forward modelling the data.

To decide on the best model for the Galaxy and the best values for its parameters one would ideally use the Bayesian framework (to decide between models, see the contribution of Trotta in this volume) combined with the maximum likelihood technique (to infer the model parameters). However, given the variety of possible Galaxy models and the complications of any particular model, which will surely have a large number of parameters, it will be very challenging to construct the priors or the likelihood functions and their derivatives (needed for their maximization). Assuming the likelihood function could be constructed we are still faced with the problem of sampling a very high-dimensional function in order to optimize the Galaxy model parameters. One way to simplify this problem is to build Galaxy models from which the probability density functions of observed quantities can be computed. The latter can then be compared to the actual data. The challenges here are the comparison of predicted and observed distributions of observables for very large amounts of data and again the exploration of a very high dimensional model space in order to find the optimum parameter values. In either case the results should be provided as a probability distributions over the model space and over the parameters of specific models. We should keep in mind that not all aspects of the Galaxy model will be uniquely determined.

4 Maximizing the science return from Gaia

As is clear from the mission capabilities described in section 1, Gaia will provide an unprecedented stereoscopic map of the solar system, the Milky Way and the nearby universe. The catalogue will contain over 1 billion stars, $\sim 300\,000$ solar system objects, millions of galaxies, $\sim 500\,000$ quasars and thousands of exoplanets. For all these objects accurate astrometry, photometry, and (for a subset) spectroscopy will be available as 'basic' data. In addition the classification, variability characterization, and astrophysical parameters of each object will be provided. When this catalogue is 'finished' around 2020 and combined with other large sky surveys it will become *the* astronomical data resource for decades thereafter, representing a tremendous discovery potential.

However as can be appreciated from the example of the modelling of our own Galaxy, maximizing the science return from Gaia is not straightforward. The true potential of the Gaia data can only be unlocked if we take an ambitious and innovative approach to data publication and access, including the provision of advanced data analysis tools. I discuss below a number of approaches that we should attempt to incorporate in the publication of the Gaia catalogue. These are at the same time areas in which further astrostatistics research is needed.

4.1 Enable hypothesis testing against the raw data

As argued in [13] all (modern) astronomical surveys produce digital intensity measurements and the most precise way to perform hypothesis testing is to forward model the raw image pixels. Any model that can explain the raw data in this way is a good model and will be constrained by every image pixel that it can generate[1]. The standard practice however is to provide a catalogue in which the raw data has been reduced to a set of standard observables, with all the 'nuisance parameters' (i.e. calibrations) already removed. The catalogue thus contains our best knowledge about the data at some point in time but with the implication that the choices about explaining the data have already been made for the catalogue user (for example, is a particular source a binary or not?). This means that hypothesis testing is severely limited by the choices made by the catalogue producers (cf. [14]).

Now, hypothesis testing against the raw data will by no means be an easy undertaking. For one, tests against the raw data require models that can also explain the calibration parameters. This is because the raw data are 'sky+telescope'. Recalibrating the data will only very rarely be undertaken so catalogue users should be offered the possibility of hypothesis testing against results from which the calibration parameters are marginalized out (in order to correctly approximate testing

[1] I should remark here that we will be interested of course in non-trivial models to explain the observations. For single stars the model 'all stars move through space at constant velocity on straight lines' will provide a good explanation of the data but it is of course not an interesting model. It does not, to name just one problem, provide us a stable Galaxy model

against the raw data). In [13] three increasingly ambitious proposals to enable hypothesis testing against the raw data are outlined:

1. Present the catalogue entries in a way that allows users to test alternative proposals for these entries (say the astrometric parameters of a star) by evaluating the resulting difference in the likelihood of the data given the model almost as if this were done against the raw image pixels. The likelihoods involved should be those for which the instrument calibrations have been marginalised out. The catalogue entries and their associated uncertainties in this proposal then become the parameters of an approximate image-level likelihood function.
2. Produce not one catalogue but many different versions that sample a posterior probability distribution of catalogues given the data. This proposal implies that there would be K versions of the Gaia catalogue that would represent a sampling from the posterior probability density function in 'catalogue space'. This approach has the advantage over (1) that star to star covariances (see contribution by Holl in this volume) can be accounted for. Any experiment or measurement is then carried out on all K samples and the resulting uncertainty then reflects the uncertainties in the primary catalogue. To properly account for all uncertainties it is important that the K catalogues should not just represent a sampling over astrophysical parameters but also over calibration parameters.
3. The previous proposal presents the problem that some of the K catalogue versions may have different complexities (a source is a binary in one catalogue but not in another). This is handled by the most extreme proposal in [13] which is to publish the full likelihood function itself. The idea is to provide the machinery that allows a catalogue user to submit a different version of the primary catalogue. The alternative version would then be evaluated by generating the predicted raw pixels corresponding to the modified catalogue and returning the difference in likelihood between the alternative and primary catalogue. Changing the calibration parameters should be allowed as well as marginalising over these.

Proposal (1) has been worked out for the SDSS-III BOSS survey (see [15]) and is already close to the way the Gaia data are currently planned to be published so it should thus be possible to implement. It will require thinking on how to do this transparently for the great variety of catalogue entries that differ a lot in their 'distance' to the raw data (for example, the magnitude of a star being more closely related to the image pixels than an estimate of its metallicity). Proposal (2) has actually been discussed before in the Gaia community and the question raised was 'how large should K be?'. One option suggested in [13] is to take the number of times a source is observed by Gaia as an order of magnitude estimate of K. In addition there is the question of how to perform the K-sampling. Both issues should be addressed through research and could potentially be tested on existing catalogues, or on the Gaia catalogue while it is built up over the course of the mission lifetime. A practical approach to partly implementing the concept of K-sampling is given in the contribution by Holl in this volume. He discusses how the star-to-star correlations in the astrometric parameters can be efficiently accounted for when averaging quantities which is equivalent to averaging over the $K+1$ catalogues. The third proposal

is clearly the most ambitious and will require a lot of research into how such an interface to the likelihood function can be practically realized and maintained (a lot of computing power will be involved). Possibly the only way to make the complicated Gaia likelihood function available is through publication of the processing software that was used (parts of which are based on forward modelling). However a major change in the attitude of users toward a 'catalogue' is required. In particular a major investment of time, effort, and computational resources will be required from the user.

The above proposals range from ambitious to possibly insane but I strongly believe they are worth considering seriously. The Gaia mission is unlikely to be surpassed for many decades to come so we will have to get the best out of what we have in hand. Getting the best out of the Gaia data will also benefit a lot from the following more modest proposals for the Gaia catalogue publication, which are in addition a prerequisite for ultimately enabling hypothesis testing against the raw data.

4.2 Preserve raw data, calibration data, and processing software

The effort described in [16] shows how better insights into the attitude modelling for the Hipparcos mission, combined with present-day computing power, enabled a higher quality re-processing of the entire Hipparcos data set. The resulting new version of the Hipparcos catalogue features very much reduced error correlations and improved astrometric accuracies (by up to a factor of 4) for the bright stars. This is the best illustration of the fact that the raw Gaia data, all the calibration data, and — very important! — the processing software, should be stored such that they are permanently accessible and readable, just as the catalogue itself will be. The raw data and calibration data (not all of which double as science data) are obviously needed for the kind of hypothesis testing advocated above. The availability of the processing software is the only practical way of allowing for the exploration of alternative calibration parameters.

The research question here is one of data curation. How do we store the raw data together with the processing software such that these are permanently accessible and *readable*? How do we make the processing software available in a way that facilitates experimenting with alternative calibrations of the science data?

4.3 Facilitate (re-) processing of the (raw) data

Already in the case of Hipparcos there are numerous examples of the re-processing of the data, notably to improve the astrometry of binaries and very red giant stars (see references in [17]). The re-processing was based on the so-called intermediate data that was published along with the Hipparcos Catalogue. The intermediate data are residuals of the observables (almost the raw data) with respect to the pri-

mary astrometric solution and the derivatives of these observables with respect to the astrometric parameters. Other example uses of the intermediate data, relevant also to Gaia, include the re-processing of intermediate data for groups of stars in order to derive a common radial velocity or parallax, the re-processing of data for objects that are discovered or confirmed to be binaries following a data release, or the re-determination of astrophysical parameters for stars following future improvements in stellar atmosphere modelling. In principle also for Gaia the re-processing of *all* the raw data might be warranted at some point in the future. In addition to the re-processing of the data the Gaia archive should also facilitate very complex operations on large chunks of the catalogue (say an all-sky search for stellar streams). Both these aims and the goal of hypothesis testing against the raw image pixels may be best served by implementing the idea of 'bringing the processing to the data' by offering users a virtual machine at the data centre hosting the Gaia archive. On this machine one could code whatever analysis or processing algorithm is called for and run it in a way specified by the user.

We will have to research the best way to present, communicate, and facilitate the use of intermediate data or raw data. Bringing the processing to the data is in principle already possible but will in practice not be trivial to implement. Partnering with private industry should be explored.

4.4 Make the catalogue and archive 'live'

A concept closely related to the previous item is that of making the Gaia data archive a 'living entity'. By this I mean that it should be possible to incorporate new information into the catalogue. Examples are complementary ground-based spectroscopy, updated classifications or parametrizations of stars based on independent information, better distance estimates for faint stars (e.g., photometric distance indicators calibrated on stars with precise parallaxes), etc. In addition the Gaia archive should seamlessly integrate with other large sky surveys including ones not foreseen at the time of the Gaia data publication. As an example, it should be possible to query the catalogue for sources brighter and fainter than the $G = 20$ survey limit of Gaia, where behind the scenes the work is done to combine Gaia and other sky surveys. One reason to do this is that the survey data from, for example, LSST is expected to form a smooth continuation of Gaia in terms of depth and accuracy as illustrated in [18].

The questions to investigate here are: how do we incorporate new information into the Gaia catalogue in a controlled manner? This means vetting of the new information, tracing the history of the information related to a source as well as the history of source classifications and parametrizations, and making the new information available in a non-confusing manner. How do we incorporate the new information as priors for the hypothesis testing against the image pixels? How do we transparently provide the combination of Gaia and other surveys, in particular searches across the different surveys?

4.5 Other issues

There are plenty of issues related to the Gaia data publication that have not been addressed above. One of them is the idea to provide, as for the Sloan Digital Sky Survey, early and frequent data releases. The arguments in favour thereof can be found in [19] and I will not say more about it here.

Further 'blue sky' thinking on the Gaia archive and future archives in general was summarized recently by William O'Mullane [20] at the request of ESA. In his report the idea of 'bringing the processing to the data' is discussed in terms of virtualization. What is not discussed in this contribution but is raised in [20] is the question of handling and visualizing the complex Gaia data. The dimensionality of the data is high (with about 10 phase space and astrophysical parameters describing each source) which makes it very challenging to interactively look for structures in the data. There is much scope here for research into data display technology and software and for investigating how to get around the 'curse of dimensionality' for algorithms that attempt classification and parametrization on high-dimensional data.

5 Future proofing the Gaia archive

The Gaia data archive in combination with other existing and future sky surveys will be the prime resource of astronomical data for decades to come as an improved Gaia mission or even a repeat of Gaia is unlikely anytime soon. The archive should therefore be 'future proof'. This not only means preserving accessibility and readability of the archive, but also not limiting the archive setup by what we imagine is possible today. Rather we should strive at publishing the Gaia data with future possibilities in mind so that one day we may indeed be able to extract the maximum possible science through hypothesis testing against the only quantities that will not change, the raw Gaia image pixels.

References

1. M.A.C. Perryman, L. Lindegren, J. Kovalevsky, E. Hoeg, U. Bastian, P.L. Bernacca, M. Crézé, F. Donati, M. Grenon, F. van Leeuwen, H. van der Marel, F. Mignard, C.A. Murray, R.S. Le Poole, H. Schrijver, C. Turon, F. Arenou, M. Froeschlé, C.S. Petersen, Astron. & Astroph. **323**, L49 (1997)
2. L. Lindegren, C. Babusiaux, C. Bailer-Jones, U. Bastian, A.G.A. Brown, M. Cropper, E. Høg, C. Jordi, D. Katz, F. van Leeuwen, X. Luri, F. Mignard, J.H.J. de Bruijne, T. Prusti, in *IAU Symposium*, *IAU Symposium*, vol. 248, ed. by W. J. Jin, I. Platais, & M. A. C. Perryman (2008), *IAU Symposium*, vol. 248, pp. 217–223. DOI 10.1017/S1743921308019133
3. C. Turon, K.S. O'Flaherty, M.A.C. Perryman (eds.). *The Three-Dimensional Universe with Gaia*, *ESA Special Publication*, vol. 576 (2005)
4. C. Turon, F. Meynadier, F. Arenou (eds.). *Gaia: At the Frontiers of Astrometry*, *EAS Publications Series*, vol. 45 (EDP Sciences, 2011)

5. A.G.A. Brown, F. Arenou, F. van Leeuwen, L. Lindegren, X. Luri, in *Hipparcos - Venice '97*, *ESA Special Publication*, vol. 402 (1997), *ESA Special Publication*, vol. 402, pp. 63–68
6. J. Binney, in *The Three-Dimensional Universe with Gaia*, *ESA Special Publication*, vol. 576, ed. by C. Turon, K. S. O'Flaherty, & M. A. C. Perryman (ESA, 2005), *ESA Special Publication*, vol. 576, pp. 89–+
7. D. Pfenniger, in *Gaia: At the Frontiers of Astrometry*, *EAS Publications Series*, vol. 45, ed. by C. Turon, F. Meynadier, & F. Arenou (EDP Sciences, 2011), *EAS Publications Series*, vol. 45, pp. 287–292. DOI 10.1051/eas/1045048
8. J. Binney, ArXiv e-prints 1104.2839 (2011)
9. W. Dehnen, J.J. Binney, MNRAS **298**, 387 (1998). DOI 10.1046/j.1365-8711.1998.01600.x
10. J. Bovy, D.W. Hogg, S.T. Roweis, ApJ **700**, 1794 (2009). DOI 10.1088/0004-637X/700/2/1794
11. A.G.A. Brown, H.M. Velázquez, L.A. Aguilar, MNRAS **359**, 1287 (2005). DOI 10.1111/j.1365-2966.2005.09013.x
12. H. Smith, MNRAS **338**, 891 (2003). DOI 10.1046/j.1365-8711.2003.06167.x
13. D.W. Hogg, D. Lang, in *Gaia: At the Frontiers of Astrometry*, *EAS Publications Series*, vol. 45, ed. by C. Turon, F. Meynadier, & F. Arenou (EDP Sciences, 2011), *EAS Publications Series*, vol. 45, pp. 351–358. DOI 10.1051/eas/1045059
14. D.W. Hogg, D. Lang, in *Classification and Discovery in Large Astronomical Surveys*, *American Institute of Physics Conference Series*, vol. 1082, ed. by C. A. L. Bailer-Jones (AIP, 2008), *American Institute of Physics Conference Series*, vol. 1082, pp. 331–338. DOI 10.1063/1.3059072
15. A.S. Bolton, D.J. Schlegel, PASP **122**, 248 (2010). DOI 10.1086/651008
16. F. van Leeuwen, *Hipparcos, the New Reduction of the Raw Data*, *Astrophysics and Space Science Library*, vol. 350 (2007)
17. M. Perryman, *Astronomical Applications of Astrometry: Ten Years of Exploitation of the Hipparcos Satellite Data* (Cambridge University Press, 2009)
18. M. Jurić, Ž. Ivezić, in *Gaia: At the Frontiers of Astrometry*, *EAS Publications Series*, vol. 45, ed. by C. Turon, F. Meynadier, & F. Arenou (EDP Sciences, 2011), *EAS Publications Series*, vol. 45, pp. 281–286. DOI 10.1051/eas/1045047
19. A.G.A. Brown, in *Gaia: At the Frontiers of Astrometry*, *EAS Publications Series*, vol. 45, ed. by C. Turon, F. Meynadier, & F. Arenou (EDP Sciences, 2011), *EAS Publications Series*, vol. 45, pp. 365–370. DOI 10.1051/eas/1045061
20. W. O'Mullane, Blue skies and clouds, archives of the future. Tech. Rep. GAIA-TN-PL-ESAC-WOM-057-01 (2011)