Gaia DR2 in the making

Lammers, U.; Brown, A.G.A.; Gracia, G.; Ballester, P.; Ibsen, J.; Solar, M.; Shortridge, K.

Astronomical Data Analysis Software and Systems XXVII
ASP Conference Series, Vol. 522
Pascal Ballester, Jorge Ibsen, Mauricio Solar, and Keith Shortridge, eds.

Gaia DR2 in the Making

Uwe Lammers,[1] Anthony Brown,[2] and Gonzalo Gracia[3]

[1]*European Space Astronomy Centre of ESA, Madrid, Spain;* `uwe.lammers@esa.int`

[2]*Leiden Observatory, Leiden, The Netherlands;* `brown@strw.leidenuniv.nl`

[3]*Gaia Project Office, Madrid, Spain;* `ggracia@sciops.esa.int`

Abstract. Gaia is ESA's space astrometry mission launched on 19 Dec 2013 into a Lissajous orbit around L2. Its main objective is to perform a stellar census of the 1000 million brightest objects in our galaxy (completeness to magnitude 20) from which an astrometric catalog of micro-arcsec level accuracy will be constructed.

Through various talks and posters (e.g. Lammers et al. 2010; Lammers & Lindegren 2014; Lammers & O'Mullane 2015), ADASS participants have been kept up-to-date over the years on various aspects of the mission from development, launch, and the first operational years to the first public data release (DR1) in September last year (Gaia Collaboration et al. 2016) The processing consortium (DPAC) and ESA are currently preparing DR2 which will be significantly richer in contents and scope compared to DR1. This contribution characterizes DR2 and outlines some of the challenges and Lessons Learned in producing a catalogue that is unique and expected to be a treasure trove for astronomical research over many years to come.

1. Introduction

Gaia is unique in many ways. From the scientific viewpoint there is little doubt now that its data will have a profound impact on many areas of astronomy and is likely to revolutionize our understanding of the Milky Way. In terms of data volumes, estimates for the total size of the final catalogue stand since years at around 1 PB which will be dwarfed by upcoming ground- and space-based missions such as LSST, SKA, or Euclid, expected to generate one or two orders of magnitude more data. However, with > 850 Billion astrometric-, > 170 Billion photometric-, and > 16 Billion spectroscopic measurements in Sept 2017, the Gaia data still represents the largest homogeneous astronomical dataset in existence at present. A first data release (DR1) was made in September 2016. DR2 will be released in April 2018 and for the first time will include results from cyclic processing activities in five of the six DPCs (Data Processing Centres) of DPAC (Data Processing and Analysis Consortium). The consortium structure with its 9 CUs (Coordination, Units), and six 6 DPCs, as well as the principles of the overall cyclic processing scheme has been described elsewhere and will not be repeated here. We rather present in the following an "appetizer" for DR2 and discuss some of the challenges and lessons learned so far from the production of the release.

2. From DR1 to DR2

DR2 is a very significant step forward, both, considering object numbers, types of products as well as scientific quality of the data (see Table 1). The "scientific core" of

Table 1. Comparison between DR1/DR2 and top-level DR2 characteristics

DR1	DR2
14 month observation data	22 months observation data
Parallaxes + Proper Motions by combining Gaia+Hipparcos	Stand-alone Gaia astrometric solution
Limited/incomplete calibrations leading to significant systematics	Much improved calibrations, much reduced systematics
G-band only	G, GBP, GRP pass-bands
Small set of variable-star light curves	Much expanded variable star catalogue
	Radial velocities
	Solar System Objects
DR2 Data Product	**Expected number of objects**
(α, δ), G, $G_{\rm BP}$, $G_{\rm RP}$	> 1.7 Billion
Parallaxes + Proper Motions	> 1.3 Billion
Median $v_{\rm rad}$ at $G_{\rm RVS} < 12$	≈7 Million
Estimates of $T_{\rm eff}$, $A_{\rm V}$, radii, luminosities for subset	Sources at $G < 17$
Photometric data for a variety of variable star types, all-sky RR Lyrae survey	≈ 200.000 RR Lyrae, ≈ 250.000 LPVs
Epoch astrometry for a pre-selected list of known asteroids	≈ 9000

DR1 were the ≈ 2 Million parallaxes and Proper Motions whose determination with merely 14 months of mission data was possible by combining the Gaia measurements with Tycho-2/Hipparcos data from some 25 years earlier using a method called TGAS (Michalik et al. 2015). DR2 will be based on Gaia data alone and give parallaxes and proper motions for more than 1.3 Billion objects. Despite its limitations, DR1 has resulted in more than 200 scientific publication in refereed journals until today which can be regarded as a forecast to the scientific harvest that DR2 will provide.

As an example of the data quality that can be expected in DR2 Fig. 1 shows the

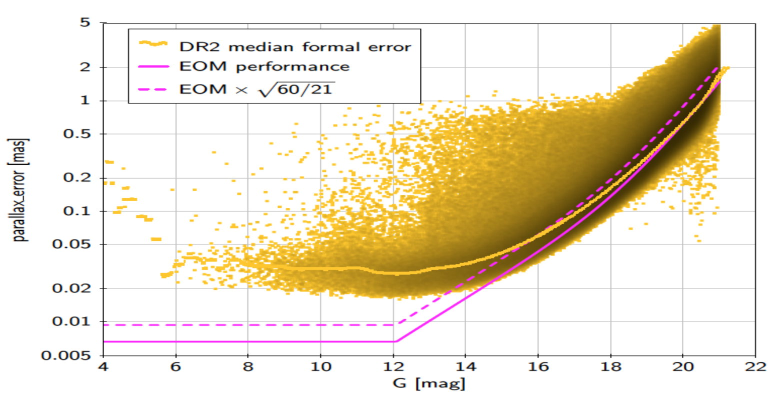

Figure 1. Formal parallax error versus magnitude expected for Gaia DR2

formal error in parallax as a function of stellar magnitude. The orange curve shows the actual median formal error in DR2, the solid blue line is the predicted end-of-mission

(EOM) performance with the full 60 months mission data, and the dashed blue line the EOM performance scaled down to the 21 months (the astrometric solution used only 21 of the available 22 months) of data in DR2. It can be seen that for stars fainter than $G \approx 16$ the DR2 parallaxes will be better than predicted and as of $G > 20$ even already now surpass the EOM performance. At the bright end, current values are about a factor 3 above (worse than) the predictions. This is mainly caused by a still imperfect instrument calibration which is the dominating element limiting the parallax precision for bright stars. The documentation accompanying DR2 will present and discuss this and more in detail. The poster presented selected "previews" for each of the product categories in DR2 (astrometry, photometry, radial velocities, astrophysical parameters, variable stars, solar system objects). To use the words of a leading DPAC scientist: "DR2 will not be perfect but it will be amazing"!

3. DR2 production challenges

Below is a subjective, and limited list of the challenges that were met during the production of DR2:

- The distributed nature of the Gaia processing and interdependencies between data (e.g. astrometry needs photometry and vice versa) necessitates that the entire complex system has to be iterated to achieve the targeted quality of the mission products. Early plans foresaw a number of such iterations at least between the major systems but owing to practical constrains (mainly time) only a single one could be carried out. This circumstance made some activities very complex such as determining a good astrometric calibration model which eliminates systematic errors to a level commensurate with the accuracy goals of DR2.

- As a side-effect of gradually improving accuracies, unexpected "features" in the data started to emerge which were previously not standing out from an overall high level of noise. Understanding and eliminating such features as part of the processing took time and (unplanned) effort. A concrete examples are the so-called micro-clanks of the payload (see DR2 documentation for details) which cause disturbances of the satellite's attitude.

- Detailed data accounting was initially not seriously considered as a possible area of concern. By this we understand the need to keep track of object numbers (stars, observations) at the input and output stages of the various distributed processes and ensuring that they are consistent. If this is not the case, the reason for discrepancies need to be understood which takes time. In a complex system such like ours data can get "lost" during transfers, storage, the actual processing (e.g. "for this raw observation there is no valid calibration") or a number of other reasons.

- As said above, future missions will "out-perform" Gaia in data volumes. However, the DR2 datasets are TBs to 10s of TBs in size and their transfers, storage, ingestion, and processing is certainly a challenge with todays computer systems.

- In general, the distributed processing of the data in 6 physically different locations is demanding in terms of transfers, coordination, scheduling, and interfacing.

- A number of software systems were run operationally only for the first time during the DR2 production. Lack of software maturity (bugs, poor performance) was then apparent in some cases despite prior testing that had been carried out. This often lead to the need of operational re-runs after issues had been fixed for which not always sufficient time had been allocated in the schedule.
- In light of the above difficulties (and many more), having only 20 months between DR1 and DR2 caused large schedule pressure for all involved groups.

4. Lessons Learned

The following is a likewise subjective and certainly incomplete and provisional list of the lessons learned during the production of DR2:

- Every system architect will have come across the temptation to declare certain conditions in data processing as impossible-to-happen ("This input value can never ever become 0"). In our experience now, with enough data (10^{12} observations in the case of Gaia), almost all conceivable anomalous conditions *will* occur! Defensive programming and planning is needed to prevent having to deal with unexpected failures caused by overly optimistic or simplistic assumptions.
- A large consortium like DPAC is not self-organizing and needs a coordinating body to function properly. For us this is the Gaia Project Office which, broadly speaking, is in charge of ensuring a proper communication flow between relevant actors, schedule devising and monitoring, and interface control.
- No complex data processing system is ready at launch — development needs to continue all through the operational phase of a mission.
- Testing systems with nominal simulation data is not sufficient — validation with flight data is needed to reach real operational maturity.
- Last but certainly not least we want to mention the importance of having dedicated, and motivated people in the team. This is beyond any doubt one of the key aspects of bringing a large project like Gaia to success.

References

Gaia Collaboration, Brown, A. G. A., Vallenari, A., Prusti, T., de Bruijne, J. H. J., Mignard, F., Drimmel, R., Babusiaux, C., Bailer-Jones, C. A. L., Bastian, U., & et al. 2016, A&A, 595, A2. 1609.04172

Lammers, U., & Lindegren, L. 2014, in Astronomical Data Analysis Software and Systems XXIII, edited by N. Manset, & P. Forshay, vol. 485 of Astronomical Society of the Pacific Conference Series, 491

Lammers, U., Lindegren, L., Bombrun, A., O'Mullane, W., & Hobbs, D. 2010, in Astronomical Data Analysis Software and Systems XIX, edited by Y. Mizumoto, K.-I. Morita, & M. Ohishi, vol. 434 of Astronomical Society of the Pacific Conference Series, 309

Lammers, U., & O'Mullane, W. 2015, in Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV), edited by A. R. Taylor, & E. Rosolowsky, vol. 495 of Astronomical Society of the Pacific Conference Series, 187

Michalik, D., Lindegren, L., & Hobbs, D. 2015, A&A, 574, A115. 1412.8770