



Source Identifiers — Assignment and Usage throughout DPAC

prepared by: U. Bastian
affiliation : ARI Heidelberg
approved by: to be approved by DPACE
reference: GAIA-CD-TN-ARI-BAS-020-01
issue: D
revision: 0
date: 06 Jun 2007
status: Draft

Abstract

A proposal is made for the assignment and usage of source identifiers throughout DPAC. Although the majority of all source identifiers will be assigned by CU3 and CU4, the scheme must be agreed across all CUs. The scheme described here must therefore be approved by the DPACE, after appropriate discussion and iteration.

Contents

1	Introduction	3
2	Format and structure of the source identifiers	3
2.1	HEALPix index and running number	3
2.2	HEALPix, DPAC version	4
2.2.1	Stars and the like	4
2.2.2	Practical usage of HEALPix	5
2.2.3	Solar-system objects	5
3	Assignment of source identifiers: the proposed scheme	5
3.1	Processes assigning source identifiers	5
3.2	Basic Principles	6
3.3	IDT cross-matching	7
3.4	Moving images and solar-system objects in IDT	7
3.5	Moving images and SSOs in CU4	8
3.6	Difficult cases in the IDT cross-matching	8
3.7	The IDU cross-matching and possible conflicts	9
4	Some additional details	10
4.1	The Initial Gaia Source List and the Initial SSO List	10
4.2	Unresolved multiple objects	11
4.3	Secondary sources from 2-d imaging	12

1 Introduction

The format and structure of the Gaia source identifiers to be used throughout DPAC were defined in the technical note “Proposal for the object numbering scheme” by F. de Angeli, F. van Leeuwen, J. Hoar, W. OMullane, GAIA-C1-MN-IOA-FDA-002-2, which was formally approved by the DPAC Executive on its second meeting. However, that document does in no way define the actual assignment, administration and usage of source identifiers for specific sources.

A corresponding scheme is proposed in the present document. Although the majority of all source identifiers will be assigned by CU3 and CU4 processes, the scheme must be agreed across all CUs. It must therefore be approved by the DPAC, after appropriate discussion and iteration.

The scheme for the assignment of source identifiers was originally drafted in a big flood of emails on Feb 5–7, 2007 among DPAC members and a few other people. The results of that email discussion was first presented in a talk by U. Bastian at the Dresden CU3 meeting in March 2007 (the Powerpoint presentation can be found via the meeting’s Wiki page, <http://www.rssd.esa.int/SA-general/Projects/GAIA/wiki/index.php?title=CU3:Core.Processing:Meetings:CU3M2>, or directly on the DPAC svn document repository, http://gaia.esac.esa.int/dpacsvn/DPAC/meetings/CU3/CU3-02-Dresden-Mar-07/bastian_SourceIds.ppt).

The present document contains only small additions to the scheme of February/March 2007.

The plan of the document is as follows: Section 2 briefly repeats the agreed basic format (from the already mentioned technical note GAIA-C1-MN-IOA-FDA-002-2) and its mathematical background. Section 3 contains the proposed scheme for the assignment, administration and usage of source identifiers for specific sources. Section 4 discusses details following from the proposed scheme which were not treated in the Feb 5–7 email discussions.

2 Format and structure of the source identifiers

2.1 HEALPix index and running number

In short, the source identifiers used by DPAC consists of a 64-bit integer, composed of two 32-bit integers:

- a HEALPix sky pixel (upper 32 bits), in the following called *index number*.
- a sequence number within the HEALPix pixel (lower 32 bits), in the following

called *running number*.

The underlying idea of labelling celestial objects by a sky pixel and a running number within that pixel is known from the Hubble's GSC and other star catalogues. HEALPix means *Hierarchical Equal-Area iso-Latitude Pixelisation* (of a sphere). It is an alternative to older systems like *HTM*, *Spherical Cube* etc., with some favourable mathematical and computational properties. The HEALPix scheme has been adopted by the WMAP and Planck projects, and now by DPAC.

More complete information on HEALPix can be found on the HEALPix homepage at JPL, see <http://healpix.jpl.nasa.gov/index.shtml> including many references, in the original paper describing it (*HEALPix - a Framework for High Resolution Discretization, and Fast Analysis of Data Distributed on the Sphere* by K.M. Gorski, E. Hivon, A.J. Banday, et al., 2005, ApJ 622, 759, previously appeared as astro-ph/0409513), or in the above-mentioned technical note GAIA-C1-MN-IOA-FDA-002-2.

2.2 HEALPix, DPAC version

2.2.1 Stars and the like

There is a small number of options to be chosen in the practical usage of the HEALPix system. For the DPAC source identifier application the choices are:

- Coordinate system: Equatorial, ICRS
- Level (fineness) of division: $N_{\text{side}} = 4096$, i.e. 6 hierarchical subdivision steps
- Pixel numbering option: “nested scheme” (i.e. not the alternative “ring scheme”)

This leads to the following practical properties of the DPAC HEALPix division:

- total number of sky pixels: $N_{\text{pixel}} \simeq 200$ million, i.e. the index numbers easily fit into a 32-bit integer
- index numbers = 1,2,3,..., $12 \cdot 4096 \cdot 4096$
- size of the pixels: about 0.7 square arcmin
- mean number of stars per pixel: about 5*/px (corresponding to 25000 */sq deg)
- Baades window: about 1000*/px (corresponding to a few million */sq deg)

The last of the items above implies that ample sufficient running numbers are available within each sky pixel, i.e. for each index number.

2.2.2 Practical usage of HEALPix

All necessary routines for the practical usage of the HEALPix system are available in the Java toolbox package GaiaTools.

2.2.3 Solar-system objects

Since they move around on the sky, solar-system objects (abbreviated SSOs in the following) cannot be assigned to a unique HEALPix sky pixel, . Thus they are given the following pseudo-HEALPix identifiers:

- index number = -1
- running number = 1,2,3,...

3 Assignment of source identifiers: the proposed scheme

The assignment of source identifiers will mainly be done by CU3 and CU4 processes, but the scheme must in the end be agreed across all CUs.

3.1 Processes assigning source identifiers

Source identifiers are primarily created and assigned to celestial sources by the following processes:

a) During the mission, more precisely during the main Gaia data processing:

- IDT cross-matching, CU3
- IDU cross-matching, CU3
- SSO matching and orbit fitting, CU4

b) Before the start of the mission:

- Preparation of the Initial Gaia Source List (IGSL), CU3

- Preparation of the Initial SSO List, CU4

Assignment of source identifiers on a smaller scale, and on a kind of secondary level, will be done by the following processes:

- double-star treatment (CU4)
- 2-d imaging (CU5)
- non-resolved multiple astrophysical objects (CU4, CU6, CU6, CU7, CU8, see Section 4.2)

3.2 Basic Principles

1. IDT and IDU cross-matching will assign source identifiers to images (SM detections), to the best of their knowledge available at runtime. As this knowledge evolves in the course of the mission and data processing, those assignments may change.
2. Both CU3/IDU and CU4/SSO must be aware of the possibility that detections belonging to the same source may initially come from IDT with different source identifiers.
3. Conversely, detections originally assigned to one source identifier may later turn out to have been misidentified, or to belong to a composite source (or composite image), thus separating into different source identifiers.
4. In consequence, both the merging of sources and the splitting of sources must be possible.
5. A source identifier, once created, will never be modified or deleted.
6. A merger of sources will create an entirely new source identifier, with a track table keeping record of the parent source identifiers.
7. Analogously, a source split will create two (or more) new source identifiers, again with a track table keeping record of what has happened.
8. In particular, the index number for a given source will never change, even if some position update would shift this source to a neighbouring HEALPix on the sky. Note that such position updates will unavoidably come about, both for technical reasons (errors in the originally used attitude, calibration and SM centroid position) and for astronomical reasons (proper motion and parallax).

3.3 IDT cross-matching

To every SM detection — or more technically, to every Astro star packet in the telemetry data stream — the IDT cross-matching will try to find a matching source in the most recent stellar source list, taking the uncertainty of the source list, attitude, calibration etc. into account. Depending on the outcome, the IDT cross-matching will take the following actions:

- In case of a match: assign existing source identifier to the SM detection.
- In case of non-match: create a new source, create a new source identifier (with index number according to detection position and running number as available), and assign this new source and source identifier to the detection.

IDT cross-matching will *NOT* try to find a matching source in the most recent SSO list, but assign ordinary source identifiers (index number > 0).

IDT cross-matching will *NOT* consider the “moving image” flags which may have been set either on board or during previous IDT processing steps.

3.4 Moving images and solar-system objects in IDT

The last two items in the preceding section are motivated by the fact that the IDT must by all means be kept streamlined, quick and fast. The IDT cross-matching will not cross-match with the SSO catalog because this is a heavy computational process that is not really necessary for the main purposes of the IDT.

Assigning ordinary source identifiers (index number > 0) to SSO detections means that “preliminary identifications” will be given to SSO observations, in perfect continuity with what the solar-system community has done for more than a century.

Also, the IDT cross-matching will not do anything special for “moving” images, because the motion can be spurious due to the superposition FoV1/FoV2 in combination with across-scan image motion, to cosmic rays, to source extension in combination with PSF variation, and so on. The “moving image” flag is merely telling that something unusual is going on, either on the sky or in the detector, thus

- prompting CU4/SSO to take a closer look at the detection
- warning CU3 and CU5 of possible astrometric and photometric problems

3.5 Moving images and SSOs in CU4

SSO source identifiers (index number = -1) will exclusively be assigned by CU4.

CU4 will produce a track table, keeping record of the parent provisional source identifier for each Gaia detection linked to an SSO source identifier.

The assignment of SSO source identifiers will be done “off line”, i.e. not in the framework of the daily IDT and FL processing at ESAC/SOC. Although CU4 may well decide to do a preliminary processing on a daily basis (i.e. directly after the delivery of cross-matched Astro elementaries from ESAC), the inclusion of the results into the Main Database and into the IDU cross-matching will take place in the 6-months cycle only.

Assignment of SSO source identifiers by CU4 can in principle be done using four basic methods:

- by positional match with known SSOs
- by confirmation of the sky motion from Gaia observations in an immediately following field-of-view transit
- by orbit reconstructions linking originally un-matched “moving” images from different epochs to each other, thus identifying previously unknown SSOs
- by orbit reconstructions linking originally un-matched “moving” Gaia images to equally unmatched ground-based observations in the IAU/MPC archives, thus giving orbits to previously observed but as yet unconfirmed/unsolved SSOs.

The latter two methods may also include “non-repeating” non-moving Gaia images, i.e. Gaia sources that received only one detection, although the one detected image was bright and point-like, and although the relevant patch of the sky was scanned by Gaia several times.

Unmatched “moving” images will (implicitly) be returned by CU4 to the IDU cross-matching. They might be reconsidered for a re-match with stellar sources, especially if subsequent semesters of Gaia observations should have produced more detections at the same position in the sky.

3.6 Difficult cases in the IDT cross-matching

This subsection just lists a number of peculiar cases to be expected. It does not set rules.

Spurious non-motion:

SSOs need not always produce “moving” images. At the tip of an opposition loop they can

move exactly on the line of sight towards the Gaia spacecraft. Such spuriously non-moving images will (in almost all cases) not be matched with stellar sources, and thus show up as non-repeating (as discussed above) later on. As such they might be matched with a known SSO by CU4.

High-proper-motion stars:

Stars with previously unknown high proper motion (or parallax) may initially receive several different source identifiers, if observed at time intervals of several months. This is not a new kind of problem; it is well known from the Hubble Guide Star Catalogue and other star catalogues. Such cases will be recognized and solved by the IDU cross-matching cluster analysis.

Crowds of spurious detections in bright-background regions (nebulae):

Care must be taken in the late stages of the Gaia data processing to remove these from the final catalogue. This will probably not be possible by fully automatic processes.

Others??

3.7 The IDU cross-matching and possible conflicts

In a 6-monthly cycle the IDU cross-matching will refine the creation and assignment of source identifiers originally done by the IDT cross-matching. This is possible due to:

- the larger number of observations available for each source (or rather: for each location on the sky)
- better attitude, calibration, PSF, source list etc.
- more time available for processing, making e.g. a cluster analysis possible

In the process, the IDU cross-matching will itself create new source identifiers. But at the same time the IDT cross-matching goes on running daily, also creating new source identifiers. Potential conflicts between IDT and IDU must be avoided. An analogous problem exists even *within* IDT and IDU, because both processes will probably be run in a highly parallelized way.

There are two obvious ways to avoid such conflicts:

- a) Bastian's method: IDT and IDU will have separate, predefined ranges of running numbers available (per sky pixel).
- b) O'Mullane's method: A central *Source Identifier Server* for the whole of DPAC (i.e. for all processes in all DPCs) will dynamically provide "packets" of spare source identifiers on request.

A completely different type of conflicts may arise between CU3/IDU and CU4/SSO: A particular detection might be assigned to an SSO by CU4, and to a stellar source by IDU. Such conflicts must be resolved by the MDB Integrator software, in a way which must have been agreed in detail between the relevant CUs before the actual processing starts.

There are quite a number of simple ways to resolve these. The two most obvious ones probably are:

- i) adopt either of the conflicting source identifier assignments, delete the other accordingly
- ii) keep both, and add a warning flag to both entries in the cross-match table

Note that occasionally an image may indeed belong to both. This clearly votes for the second method. That method also covers the case that an image may belong to two different stellar source identifiers — due to the superposition of the two fields of view, or due to scanning through a resolved double star at an unfavourable scan direction. In the very last run of the MDB Integrator software — producing the final match table in 2019 or 2020 — it may be appropriate to use a more sophisticated conflict resolving scheme than in the routine 6-months iterations before.

4 Some additional details

4.1 The Initial Gaia Source List and the Initial SSO List

A large number of source identifiers can (and will) be pre-existing at Gaia's launch:

- the members of the Initial Gaia Source List, derived from GSC-II etc., including the reference stars and standard stars for astrometry, photometry, radial velocities, stellar classification, attitude determination and so on
- the Initial SSO List, containing the already known SSOs from the IAU/MPC database, maybe even including the database of un-matched ground-based SSO observations

Source identifiers newly created from Gaia observations might be distinguished from these pre-existing ones by a gap in the running number,

- e.g. newly created ones starting at 10 001 or $2^{16}+1=65\,537$ for stellar sources
 - e.g. newly created ones starting at 10 000 001 or $2^{24}+1$ for SSOs
- or by corresponding flags in the respective source lists.

A few additional remarks on the Initial Gaia Source List (IGSL) may be appropriate here. The IGSL will be provided by CU3 (GWP-S-335-11000). It is intended to provide a rough pre-launch approximation to the expected Gaia sky. It will be constructed from the best available ground-based sky survey(s) before Gaia's launch, appended by special source lists like known cataclysmic variables, known QSOs etc., and cross-matched with auxiliary catalogs of reference and standard stars for

- photometric calibration
- astrometry (attitude stars, QSOs)
- radial-velocity calibration
- source classification and astrophysical parameterisation
- variability detection
- the Gaia ecliptic-poles catalogue (for commissioning and initial calibration)
- etc.

The IGSL is intended to form the zero version of the main Gaia source list, to be revised and improved in the successive MDB versions. The IGSL will be far from complete for many reasons, the single most important one probably being the low angular resolution of ground-based sky surveys compared to Gaia. Nevertheless it will be very useful, e.g. for the management of the sets of reference and standards stars just mentioned, and for the management of SSOs.

4.2 Unresolved multiple objects

Source identifiers are effectively created by and assigned to SM detections. Thus the philosophy of assigning source identifiers — as described in Section 3 — implicitly rests on the concept that two different sources generally¹ should create separate SM detections.

However, in many ways Gaia will detect that perfectly pointlike sources consist of two or more physical objects: Components of spectroscopic binaries (from radial-velocity measurements), of eclipsing binaries (from photometry), of astrometric binaries (from astrometry), of composite-spectrum stars (from astrophysical classification), etc. In many cases it will be possible to even individually characterize these physical objects.

The designation of such components of unresolved multiple objects (spatially unresolved by Gaia, that is) should — to my opinion — not be done by creating entirely new source identifiers. Instead they should be treated as “components” of the sources identified according to the scheme in Section 3. In other words: the Gaia source identifiers as defined in Section 2 should refer to “Gaia spatial-resolution items”.

The detection and classification of physical components of pointlike sources is much less secure, much more ambiguous, and much more complicated (due to the potential involvement of many CUs at the same source) than the assignment of source identifiers to SM detections. Creating

¹ Exceptions are the superposition of images from different fields of view and the superposition of double-star components in unfavourable scan directions

entirely new source identifiers would potentially lead to a very large size and complexity of the track table defined in items 6 and 7 in Section 3.2. Also, this would be in contrast to usual astronomical practice. Occasionally it would lead to inappropriately large emphasis on uncertain classification issues.

The obvious advantages of the concept of “components” must be weighed up against the addition of a third part to the source identifier for components, in addition to the index number and the running number. This third part, the *component number*, will add complexity and data volume to the source designations, but — to my opinion — this is to be preferred over the additional volume and complexity of the track table. The component number can be a `short` integer or even a `byte` in the Java sense.

4.3 Secondary sources from 2-d imaging

The concept of “components” described above should perhaps also be used for 2-d imaging components. The Tycho data reduction did exactly that in fairly analogous circumstances.

This issue has not been thought through yet. It is more complicated than the case of the multiple physical components of point-like sources, because a given celestial source can at the same time be an independent Gaia source (in the sense of Section 3) and turn up as a 2-d imaging component of a separate, neighbouring Gaia source. So the arguments of the preceding subsection do not hold as strictly in the present case.

References

to be done