Source Identifiers — Assignment and Usage throughout DPAC

prepared by: U. Bastian, J. Portell
affiliation : ARI Heidelberg
approved by: DPACE (Issues 1-4), CU3-L (Issue 5)
reference: GAIA-C3-TN-ARI-BAS-020-05
issue: 05
revision: 0
date: 2020-04-30
status: Issued

Abstract

Issue 4 documents the changes agreed at the 7^{th} CU3 meeting and implemented in GaiaTools in 2012. Issue 5 implements some minor fixes and adds a diagram. A definition is given 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 basic scheme described here was approved by the DPACE, after appropriate discussion and iteration; the changes and additions in Issue 5 were approved by the CU3-L.

Document History

Issue	Revision	Date	Author	Comment
5	0	2020-04-30	JP	Provided a bit more details on SSO SourceIDs. Mentioned JP-079 for the component numbers. Other minor revisions and corrections. Issued to Livelink.
4	1	08-May-2018	JP	1) Sect. 2.1: 3-bit DPC code (not 2-bit); no spare bit anymore; added footnote to make clear that the sourceId HEALPix is determined only at the time the source was created. 2) New subsection (2.7) with diagram. 3) New subsection (2.8) with "human" decoding. 4) Mentioned that, in practice, VO SourceIDs seem not to be used at all (end of 2.4). 5) Small clarification on sequence numbers (multiples of 128, in 2.5). 6) DPC values updated and clarified further (2.6). 7) Mentioned in 3.3 that IDT NewSources are not used for the final catalogue. 8) Mentioned in 3.4 that we may consider using the reserved DPC codes for exhausted superpixels (should never happen). 9) Bibliographic references changed from `cite` to `citell`.
4	0	18-Jun-2013	BAS	BAS-038, CU3M7, CF's presentation incorporated
3	0	12-Jun-2010	BAS	More detail on SSOs and on components; reference to BAS-033; observed flag and superseded flag for Integrated Source
2	0	12-Dec-2008	BAS	redefinition of index number for empty windows (virtual objects, see Section 2.2.4), plus a few purely editorial changes
1	1	28-Sep-2007	BAS	Feedbacks from the DPACE and others, in particular: 1) embedding of component numbers in running numbers, 2) "superpixels" for management of running numbers, 3) addition of numbering for forced windows, 4) note on inconsisteny of IGSL prototypes, 5) removal of gaps in running no. to indicate IGSL membership, 6) parallel (second) match table for SSOs
1	0	06-Jun-2007	BAS	Creation; submitted for approval

The source files for this document can be found in
`http://gaia.esac.esa.int/dpacsvn/DPAC/CU3/docs/General/BAS-020-Source-Ids/`

Contents

1 Introduction

The format and structure of the Gaia source identifiers to be used throughout DPAC were originally defined in the technical note "Proposal for the object numbering scheme" (FDA-002) 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 defined 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 was therefore approved by the DPACE, 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 DPACE 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 or more directly on the DPAC svn document repository.

The present document (since issue 1) contains additions and clarifications to the scheme of February/March 2007, collected during the DPACE approval process during summer 2007. Issue 2 corrected a conceptual error found by C. Fabricius and J. Portell in Dec. 2008. Issue 3 added more detail on the treatment of components and of solar-system objects, and put the present document in perspective to BAS-033 on the usage of the IGSL. Issue 4 documents the changes in the detailed structure of source identifier agreed between CU1 and CU3 at the 7^{th} CU3 meeting and implemented in GaiaTools in 2012. That agreement is described and discussed in the minutes (BAS-038), and in a presentation by C. Fabricius (hyperlinked in BAS-038). Issue 5 includes a diagram of the SourceID structure, the description of a human-readable format, and includes some minor fixes (such as the number of DPC bits, which were wrong in Issue 4).

The plan of the document is as follows: Sect. 2 briefly describes the agreed basic format and its mathematical background. Sect. 3 contains the proposed scheme for the assignment, administration and usage of source identifiers for specific sources. Sect. 4 discusses details following from the proposed scheme which were not treated in the Feb 5–7, 2007 email discussions.

2 Format and structure of the source identifiers

The format and handling of source identifiers is supported in GaiaTools by the Java interface `gaia.cu1.tools.util.SourceIdUtil`.

2.1 HEALPix index, running number, component number, DPC code

In short, the source identifiers used by DPAC consists of a 64-bit integer, comprising:

- a HEALPix sky pixel[1] in bits 36–63 (where lsb=1, msb=64), in the following called *index number*. By definition the smallest HEALPix index number is zero. More details are given in Section 2.2.
- a 3-bit DPC code, defined in Section 2.6, in bits 33–35 (lsb=1, msb=64)
- a 25-bit plus 7-bit sequence number within the HEALPix pixel in bits 1–32 (where lsb=1, msb=64), split into:
  - a 25-bit *running number* in bits 8–32 (lsb=1, msb=64). The running numbers are defined to be positive, i.e. never zero (except in the case of forced empty windows, see Sect. 2.4).
  - a 7-bit *component number* in bits 1–7 (lsb=1, msb=64). The component number is described in Section 2.5.

The underlying idea of labelling celestial objects by a sky pixel and a running number within that pixel is known from the Hubble Space Telescope'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.

[1] It is determined from the sky coordinates *at the time the source was created.* The HEALPix of the sourceId is *never* updated (otherwise the sourceId value would obviously change). It means that, in some cases (mainly IGSL sources), the HEALPix contained in the sourceId may be quite far (even some arcminutes) from the actual source coordinates.

2.2 HEALPix, DPAC version

Complete information on HEALPix can be found on the HEALPix homepage at JPL 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 (FDA-002).

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: Nside = 4096, i.e. 12 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: $N_{pixel} \simeq 200$ million[2], i.e. the index numbers easily fit into a 32-bit integer
- index numbers = 0,1,2,3,..., 12*4096*4096-1
- size of the pixels: about 0.7 square arcmin
- mean number of stars per pixel: about 5*/px (corresponding to 25000 */sq deg)
- Baade's window: about 1000*/px (corresponding to about 5 million */sq deg)

The last of the items above implies that only about 10-12 bits (out of the available 25 bits) would be needed if the running numbers would be used for just trivial counting per pixel.

However, such a simple scheme would run into practical difficulties (see Sect. 3.4) and would not take the planned cyclic revision of the Gaia source list into account (see Sect. 3.8).

2.2.1 Practical usage of HEALPix

All necessary routines for the practical usage of the HEALPix system are available in GaiaTools, in the Java interface `gaia.cu1.tools.util.SourceIdUtil`.

[2]Specifically: there are 201,326,592 level-12 pixels, that is, 12×4^{12}. In hexadecimal that is 0x0C000000.

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,... (i.e. positive)

The specific assignment of running numbers to SSOs is not a subject of the present document. It is defined by CU4 in FM-036, in accordance with the present document.

2.4 Forced empty windows (background windows, virtual objects)

The forced empty windows (also known as background windows or virtual objects) produced by the Gaia instrument cannot be assigned to any specific celestial source. Furthermore, they do not belong to a specific field of view (FoV), but almost[3] equally to both fields. Thus they are given the following pseudo-source identifiers:

- index number = -2 - (HEALPix index number of the window center in the preceding FoV)
- running number = +1 * (HEALPix index number of the window center in the following FoV)

In this way the rough (0.3 arcmin rms) location on the sky of both FoVs is given for the windows. The offset[4] of -2 is introduced because index number -1 is already occupied by the SSOs and index number zero is occupied by stellar sources. It is no problem that the same pair of index number and pseudo running number may rarely (if at all) be assigned to more than one such window.

Note that, as of Data Reduction Cycle 3 (or Issue 5 of this document), we do not know of any DPAC system actually assigning SourceIDs to Virtual Object windows.

[3] the word “almost” in this sentence refers to the fact that there is only one SM window for each virtual object. The transit identifier (see JP-011) of the corresponding raw and intermediate data records indicate which of the two SMs is involved

[4] note that this was different — and incorrect — in issue 1 of the present document

2.5 Component numbers

Since Gaia source identifiers are primarily assigned on the basis of SM detections (by IDT and IDU), they by necessity are defined in terms of the resolution properties of the 2-d images of the SM and of the on-board detection algorithm. In other words, a source identifier primarily denotes an SM detection item on the sky.

However, as detailed in Sect. 4.2 and Sect. 4.3, such an item may later split into several distinct astrophysical objects. For this reason the lowest 7 bits of the sequence number[5] are reserved for the 'component number'. In other words, the true source counter starts on bit 8 only. The smallest component number (in particular the one for sources which have not split into components at all) will always be zero. Including the lowest seven bits, the smallest sequence number[6] within each HEALPix will thus be 2^7=128, and different sources (not components) in a given pixel will get sequence numbers differing a multiple of 128.

In the first draft of the present document, the component number had been assigned to a separate quantity, adding volume and complexity ot the data structure. Its introduction was independently proposed by R. Smart, C. Fabricius and Torra. The choice of precisely 7 bits (rather than the perhaps more obvious choice of 8) is motivated in Appendix A.

2.6 DPC code

In order that the data processing at different DPCs can independently assign source identifiers without creating inconsistencies, a DPC code was added to the source identifier in 2011. It is a 3-bit binary code with the following values:[7]

- DPCT (IGSL) = 0
- DPCB (CU3 IDU) = 1
- DPCI (CU5) = 2
- DPCC (CU4) = 3
- DPCE (CU3 IDT) = 4
- REP (reprocessing) = 5, NA1 (reserved) = 6, NA2 (reserved) = 7

More motivation for this is given in Sections 3.4 and 3.8.

[5] Not for solar-system objects and for forced empty windows

[6] Note that we intentionally use "*sequence number*" (which is 25+7 bits). The "*running number*" refers to only the 25 bits part, which starts from 1.

[7] Implemented in GaiaTools as: `public enum gaia.cu1.tools.util.SourceIdUtil.Location` and `public char[] locationC ={'T', 'B', 'I', 'C', 'E', 'R', 'X', 'Y'};`

2.7 SourceID diagram

Fig. 1 illustrates the structure of the DPAC SourceIDs.

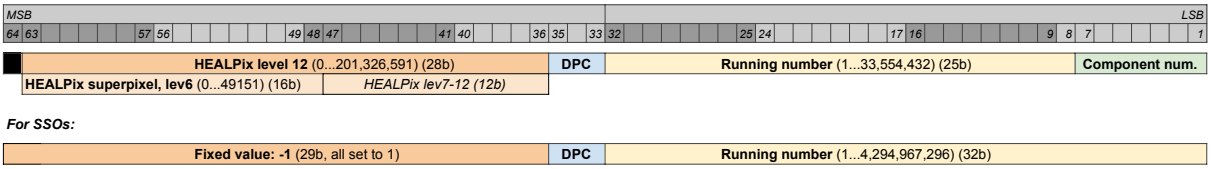

FIGURE 1: Overview of the SourceID codification.

2.8 Human-readable SourceID format

At the already mentioned CU3M7 meeting (BAS-038), a "human-readable" format of the SourceID was also proposed, to make it slightly simpler to read than a plain 64-bit number. It depends on the nature of the SourceID — that is, whether it indicates a normal source, an SSO, or a VO.

For normal sources the format is the following:

`HEALPix-DPC-number-comp`

where:

- `HEALPix` is the level-12 pixel indicated as `[N,E,S]` (for North, Equator or South) + `[0-3]` (the top-level pixel within the N/E/S region) + six-digit hexadecimal value
- `DPC` is one letter: `T`, `B`, `I`, `C` or `E`. The remaining `R`, `X` and `Y` are not used yet
- `Number` is the running number (from 00000001 to 33554432)[8]
- `Comp` is the component number (from 000 to 127)

[8] It is indicated with a decimal number (requiring 8 digits) instead of hexadecimal (requiring 7 digits). The difference is not worth using hexadecimal, and a decimal number may be slightly more useful here.

For SSOs, the format is much simpler:

`SSO-number`

where:

- `SSO` is the fixed "SSO" string
- `number` is the running number assigned by CU4, with 1 to 10 digits, in plain decimal (not hexadecimal)

This decoding is available, for example, in the on-line SourceId Decoder at `https://gaia.esac.esa.int/decoder/`. Some examples:

- Sirius b is 2947050466531873024, which is decoded as E11CC0D9-B-102434-0
- Proxima Centauri is 5853498713160606720, decoded as N22779C2-B-412104-0
- Kapteyn's Star is 4810594479417465600, decoded as N0585546-B-17462-0
- Haumea (an SSO) is -4283606216, decoded as SSO-11361080[9]

[9] As a curiosity, the running number assigned by CU4 is based on the corresponding number from the Minor Planet Centre (if available), but it does not exactly coincide. For example, Haumea is officially 136108, but CU4 uses 11361080. They use some specific rules to differentiate between known bodies, new ones, major planets, natural satellites, etc.

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, see Sect. 4.3)
- source environment analysis, also known as 2-d imaging (CU5, see Sect. 4.3)
- non-resolved multiple astrophysical objects (CU4, CU6, CU7, CU8, see Sect. 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. The associated astrometric and photometric parameters will change. The list of observations associated (via the match table) to a given source identifier will change, too.

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).

C. Fabricius and J. Torra propose to waive items 6 and 7 above in particularly trivial cases of merging and splitting: Adding just one additional observation (which for some reason was on its own with a separate source identifier) to a well-established source, or removing one observation that at closer look does not belong to the originally assigned source identifier. The former case can be trivially accomodated by the track table; the latter case with appropriate flagging.

3.3 IDT cross-matching

To every SM detection — or more technically, to every Astro star packet (SP1) 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 assign one and only one source identifier to each SM detection. If, according to the current match criteria, there is more than one match, a nearest-neigbour criterion will be used to decide. Brightness considerations will generally not be used. A star can strongly and erratically change its brightness, but not so easily its position.

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). Furthermore, IDT cross-matching will *not* consider the "moving image" flags which may have been set either on board or during previous IDT processing steps. The goal of these two decisions is to limit he complexity and computational load of IDT. One side effect is that the science alerts task (if it should be retained) will have to check for known SSOs before issuing an alert.

When starting DPAC Operations, it was finally decided that IDT NewSources would *not* be taken into account by the IDU XM, which means that IDT SourceIDs will never appear in the final Gaia catalogue. Instead, they will only be seen by the daily pipelines within DPAC.

3.4 The concept of "superpixels"

Assigning new running numbers in each of the 200 million sky pixels independently would force IDT to keep track of the next available number in 200 million instances, which would be quite a burden. This problem — which analogously exists for IDU and all other processes potentially assigning source identifiers — was discovered by J. Castañeda, who also proposed the solution given here.

The full level-12 HEALPix index number will be assigned to each individual source, but using the same series of running numbers for a large subset of the level-12 pixels. Taking advantage of the hierarchical structure of HEALPix, this means allocating only one series of running numbers in an entire level-6 pixel, i.e. common to all the 4096 level-12 pixels of each level-6 pixel. In this way IDT (etc.) will only have to keep track of the next available running number for the 49 152 level-6 pixels on the sky.

This concept of superpixels for the assignment of running numbers will eat up 12 bits of the 25 bits available for the running number. As can be seen from Appendix A, this is easily affordable.

In principle, if a shortage of bits for the running number should have arisen, the HEALPix level of the superpixels could be increased to level 7 (saving two bits and increasing the tracking of the next available running number to 200 000 instances) or even level 8. This possibility

was discussed in 2012 at the 7th CU3 meeting and was finally discarded in favour of a longer running number (25 bits instead of the original 24), see BAS-038 and JDB-075.

In the (really improbable) event of a shortage of bits, we may consider to restart the running number in that exhausted pixel and use one of the reserved DPC codes.

3.5 Moving images and SSOs in IDT

Two of the last items in section 3.3 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 several reasons: the superposition FoV1/FoV2 in combination with across-scan image motion, cosmic rays, 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

3.6 Moving images and SSOs in CU4

SSO source identifiers (index number = -1) will exclusively be assigned by CU4. That is, CU4 will set up the match table between SM detections and SSOs. This match table will be kind of parallel to the match table produced by IDT and IDU cross-matching, see last paragraph of this subsection.

In addition, 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.

Following a suggestion by F. van Leeuwen and (indirectly) A. Spagna, IDU will ignore the SSO assignments done by CU4. That is, the images will in the primary match table retain their provisional source identifiers from IDT.[10] This means that e.g. CU4 trying to link data to SSOs will not see "their" data change identifier with every semi-annual run. This also removes the need to identify SSOs in IDU, where it would be an equal burden as in IDT.[11] There is the further advantage that for each of the observations it is clear where in the sky it took place. The fact that the same observation might thus appear both in the "stellar" match table and in the SSO match table does not pose a problem. Such parallel appearance might, quite on the contrary, be taken as an easily accessible warning that any single assignment might be dubious. More details are given in Section 4.5.

3.7 Peculiar cases in the IDT cross-matching

This subsection just lists a number of special 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

[10] except if they should get matched to other observations, thus creating a real stellar source.
[11] Presently, SSO processing is kept as an option in the requirements specifications for IDU.

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. In any case they do not pose a problem for IDT.

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 cases may also appear.

3.8 Possible conflicts

3.8.1 IDT vs. IDU

In a 6- to 12-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 more detailed 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 *within* 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.

Initially it appeared that O'Mullane's method is the preferrable one. Since the scheme must run simultaneously for several CUs, CU1 was made responsible for developing and operating an envisaged central *Source Identifier Server*. However, it turned out to be impractical due to the distributed nature of the source identifier assignment. Therefore, as a compromise option combining the advantages of a) and b) above, the DPC code was introduced:

c) Each DPC runs its own independent *Source Identifier Server*, and independently keeps track of the available running numbers within each superpixel.

Note that this in effect means the extension of the running number by three more bits (the ones corresponding to the DPC identifier), and that the same combination of Healpix index and running number may occur up to seven times (one per DPC).

3.8.2 Stars versus solar-system objects

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; add a warning flag to the retained one

ii) keep both; 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.

3.9 Assignment of component numbers

As of Issue 4, the precise scheme for the assignment of component numbers is not treated in the present document. It is to be defined in a consensus between CU4 (double and multiple stars) and CU5 (2-d imaging). A future issue of the present document (or a separate document) should record it, once formed and finally agreed. Specific considerations, both on resolved and unresolved components, are given below in Sections 4.2, 4.3 and 4.4.

As of Issue 5, a separate document finally provides the necessary details about component numbers in the SourceIDs: JP-079.

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, PPMXL, UCAC-4 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 will be distinguished from these pre-existing ones by a flag, and by the cross-match tables of the IGSL. When Gaia starts to create new source identifiers, i.e. immediately at the beginning of scientific observations, the cross-matching software must be aware of the already "spent" running numbers in each superpixel (and vice versa of the still available ones). The concept of superpixels must be taken into account in the creation of the IGSL source identifiers already.

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) and 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 a kind of 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 is very useful, e.g. for the management of the sets of reference and standards stars just mentioned, and for the management of SSOs. Whether the source inventory of the final Gaia catalogue should be "grown" out of the IGSL or from scratch (i.e. out of an initially empty source list) has been under discussion for about three years. This discussion was ended by BAS-033 in January 2010.

During the development of the IGSL a number of prototypes were be delivered to DPAC. However, until the final version there was no attempt to keep the source numbering consistent. The first prototype produced in 2007 was largely based on GSC II, while the final version (produced in early 2013) was based on a larger number of sky surveys not even existing in 2007. This would have made a consistent numbering difficult (if not impossible) without any significant benefit.

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 Sect. 3 — implicitly rests on the concept that two different sources generally[12] 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) will therefore not be done by creating entirely new source identifiers. Instead they will be treated as "components" of the sources identified according to the scheme in Sect. 3. In other words: the primary Gaia source identifiers (bits 8 to 31 of the running number) as defined in Sect. 2 will 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.

[12]The exception being the superposition of double-star components in unfavourable scan directions

Creating 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 Sect. 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" clearly outweigh the small additional complexity introduced by separating the lowest 7 bits of the running number. This point is strengthened further by the considerations in the following subsection.

4.3 Secondary sources from 2-d imaging and from double-star treatment

The concept of "components" described above will also be used for 2-d imaging components and for extra sources found in the double-star treatment of disturbed images etc. 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 Sect. 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. The present case must thus be deferred to the detailed definition of the MDB Integrator, which is to be done by CU1 in consensus with CU3, CU4 and CU5.

Sketchy ideas how such a consensus might look like have been suggested at the 2007 meeting on 2-d imaging in Brussels.[13] Astrometric solutions might be adopted in a kind of hierarchical priority scheme: CU3 (5-parameter solution for single source) will be superseded by CU5 (5-parameter solutions for several components), which in turn will be superseded by CU4 (multiple-star solutions). The details of 2-d imaging work and outputs for clearly resolved sources and in crowded regions remains to be defined.

4.4 The fate of "parent" SourceIds that have acquired components

As stated in Section 3.2, a SourceId once created will never be modified or deleted. This must in particular hold for the original SourceId (with component number zero) of a source that later splits into components. Any Integrated Source List in the Main Database (produced by the MDB Integrator at the end of a DPAC processing cycle) will thus contain a "parent" or "system" record (containing information from CU3), carrying component number zero, plus one record for each component (containing information from CU4-8). This important rule was set up after careful discussion at the 3rd IDT/IDU Cross-Match Meeting at Barcelona on June 1/2, 2010.

[13] This paragraph follows an informal report by C. Fabricius

The reasons are:

- By definition, IDT and IDU never enter the component level. They only treat SM detections in their cross-matching and image parameter determination processes. These are represented by the "system" sourceId of any components that might be found later on. If the "system" sourceIds would be removed from the source list, IDT and IDU would unnecessarily and uselessly create ambiguous cross-matches for the whole set of components.
- AGIS, using the cross-match tables from IDT and IDU, can only do a single astrometric parameter adjustment for each "system" SourceId, but not for components. The astrometric parameters for the "system" SourceId are recorded in the appropriate MDB table record. Any astrometric improvements for individual components found by CU4 or CU5 will be recorded in the component records. They will logically replace the "system" parameters whenever relevant for downstream users in DPAC.
- Local Plane Coordinates used by CU4 and CU5 are computed with respect to the position given by an AGIS solution. The "system" record of the source list gives the unique and unambiguous reference for this kind of operation.

The "system" astrometric parameters from AGIS may in some cases be astronomically quite strange in nature (referring to the location and motion of some centre of light of the system), but they are the unique and well-defined reference from which to investigate the components.

In particular, IDT and IDU will use only the "system" records (those with component number zero) when setting up their internal working star catalogues from MDB Integrator inputs (and from the IGSL).

4.5 The SSO cross-match table and the fate of provisional SSO SourceIds

The rule stated in item 5 of Section 3.2, that a SourceId once created will never be modified or deleted, also holds for the provisional SourceIds given to SM detections of solar-system objects. They will always be kept in the integrated source lists, for reasons given below.

In addition, the 3rd IDT/IDU Cross-Match Meeting at Barcelona on June 1/2, 2010, after careful discussion decided that the "stellar" cross-match table (from CU3, IDT/IDU) and the SSO cross-match table (from CU4) will never be merged in the MDB.

Reasons for these decisions are:

- An SSO match created by CU4 may at a later time be retracted by CU4. In that case the original sourceId will revert to a stellar one, so to speak. Reasons for such a retraction may be a temporarily bad attitude used for the initial cross-match, an incorrect or inaccurate orbit assumed for the SSO, a grossly disturbed SM image location, or whatever.
- Even a correct SSO match may at a later stage turn out to be a superposition of an SSO and a star. Such cases will be recognized by the fact that later on, with more and more mission data accumulating, more matches with the provisional sourceId are found by the IDT/IDU cross-match processes. Removing the original sourceId and cross-match record would lead to the re-creation of the star under a new sourceId, thus potentially clouding out the superposition and adding unnecessary sourceIds to the system.
- Such superpositions will be good to know for the scientific treatment of both the star and the SSO.
- There will be quite a significant number of such cases: The total area of the cross-match circles around the (average of) 80 detections for each of the (roughly) 10^6 SSOs is of the order of $6{\cdot}10^7$ square arcsec (assuming a cross-match radius of 0.5 arcsec). On this area of the sky we have of the order of one million Gaia stars, i.e. there will be of the order of one million such cases in the Gaia data. Actual superposition of images will be about a factor of 10 less frequent, but even these will add up to the order of 100 000 cases.
- Finally, these decisions avoid unnecessary logical interfaces between CU3 and CU4 and unnecessary actions on the side of the MDB Integrator.

4.6 The fate of non-existent IGSL sources and of superseded SourceIds

Some of the sourceIds from the IGSL will never be observed, for several possible reasons: because the corresponding source does not exist, because the source is actually fainter than the IGSL indicated, because it is a variable star that happens to be always too faint when touched by Gaia, and so on. These cases will be recognized by the fact that the corresponding sourceIds will never appear in any cross-match table.

It is a task of the MDB Integrator to distinguish such sourceIds from those that have actually received Gaia observations. Only the latter can reasonably (and in fact must!) be treated by the cyclic DPAC processes. The means to do so is a Boolean flag (the “observed” flag) in the Integrated Source list delivered by the MDB Integrator for each MDB version. In the zero version of the MDB (the one at Gaia’s launch, just representing the IGSL) this flag is set to “false” for all sources. In the preparation of later MDB versions, the MDB Integrator sets this flag to “true” for all sourceIds that occur in any of the cross-match tables relevant for the MDB version presently to be released.

Although the “observed” information could in principle be produced by any user of the Integrated Source List (by making use of the relevant XM tables for the MDB version under consideration), this would be very inconvenient and would be a waste of effort and resources. Rather than doing this job several times at different DPCs and for/by different software systems, the MDB Integrator can do it in a consistent and streamlined way while compiling the Integrated Source List for the next forthcoming MDB release/version.

Even at the end of the mission there will be a lot of “false” values to this flag. The corresponding sourceIds will of course not appear in the published Gaia catalogue. Note that the status of the “observed” flag for a given sourceId may in principle revert from “true” to “false”. This happens if IDU cross-matching should retract an IDT cross-match (which may spuriously have been created due to bad attitude OGA1 or bad IDT image location).

As for sourceIds made obsolete by a merger, split or any other source list revision (by IDU or the MDB Integrator): These will not disappear from the Integrated Source table(s), but will be marked by a “superseded” flag. The setting of this flag again is an MDB Integrator task. It will be initially set to “false” for all sourceIds. It will be set to “true” by the MDB Integrator for all sourceIds that appear as the “old” sourceId in any track table record relevant for the MDB version presently to be released. Attention, this time it is the track table(s), not the XM table(s). (Note: for sources that already have “true”, the check with the track table(s) needs not be repeated.)

Again, this information could in principle be produced by any user of the Integrated Source List (by making use of the relevant track table(s) for the MDB version under consideration), but this would be very inconvenient and would be a waste of effort and resources. Rather than doing this job several times at different DPCs and for/by different software systems, the MDB Integrator can do it in a consistent and streamlined way while compiling the Integrated Source List for the next forthcoming MDB release/version.

References

[BAS-033], Bastian, U., O'Mullane, W., Portell, J., 2010, *The starting source list of IDT*, GAIA-C3-TN-ARI-BAS-033, URL http://www.rssd.esa.int/cs/livelink/open/2966089

[BAS-038], Bastian, U., de Bruijne, J., Gracia, G., et al., 2012, *Minutes of the seventh CU3 plenary meeting (CU3M7)*, GAIA-C3-MN-ARI-BAS-038, URL http://www.rssd.esa.int/cs/livelink/open/3127683

[FDA-002], De Angeli, F., van Leeuwen, F., Hoar, J., et al., 2007, *Proposal for the object numbering scheme*, GAIA-C1-MN-IOA-FDA-002, URL http://www.rssd.esa.int/cs/livelink/open/2698292

[JDB-075], de Bruijne, J., 2012, *With HST through the LMC to the Gaia SourceId*, GAIA-CD-TN-ESA-JDB-075, URL http://www.rssd.esa.int/cs/livelink/open/3125660

[FM-036], Mignard, F., 2008, *Gaia Identifiers for Solar System Objects*, GAIA-C4-TN-OCA-FM-036, URL http://www.rssd.esa.int/cs/livelink/open/2851628

[JP-079], Portell, J., 2020, *Specification of a Source Consolidator*, GAIA-PO-TN-UB-JP-079, URL http://www.rssd.esa.int/cs/livelink/open/1255425

[JP-011], Portell, J., Bastian, U., Fabricius, C., et al., 2018, *Updated definition of a unique Transit Identifier for the MDB*, GAIA-C3-TN-UB-JP-011, URL http://www.rssd.esa.int/cs/livelink/open/2811678

Appendix A: Motivation of the bit numbers

There are 25 bits available for the running numbers (but note that they are defined to be positive, i.e. the zero value is excluded). This corresponds to about 34 million available numbers per superpixel (i.e. per 0.8 square degrees), or about 40 million sources per square degree.

The practical limit for the density of sources observable by Gaia is defined to be in the range of 3 million sources per square degree. This limit will never be reached over a full square degree, however. Thus the 25 bits for each superpixel provide a generous margin to make sure that the running numbers will never be exhausted. This topic was more carefully discussed and brought to a firm conclusion by J. de Bruine in JDB-075.

A big margin is needed because running out of numbers would be a desastrous failure of the scheme. Furthermore, the continuous revision of the numbering due to mergers and splits of sources necessitates more numbers than actually observed sources. Mergers and splits will obviously affect a minority of all sources only. But even if — in case of a misfunction of the on-board detection or other unforeseeable very unfavourable circumstances — this would need, say, an *average* of 2 or 3 source identifiers per actually observed source *at the maximum star density and over a full square degree*, we would still be on the safe side.

It should be noted that other usages of running numbers, like the provisional source identifiers for SSOs and the assignment of source identifiers to spurious detections (cosmic rays, noise peaks etc.), are negligible in this context.

It should furthermore be noted that it would not be a real problem if the available number of 128 components per running number should ever be exhausted. In this case we (specifically: The Main Database Integrator software) will simply create a new running number and start the component numbering from scratch. It is clearly impossible that Gaia will ever need to describe more than 128 components within the sky region defined by a sky mapper detection, which is in the order of 0.3 arcsec. Nevertheless it is conceivable that the available range of component numbers may be exhausted in very unfavourable cases. This may in principle occur due to frequent revision of the subdivion into components and/or due to the desire for a “systematic” subdivision of a source in the case of a highly hierarchical astrophysical system.