Équipe Gaia
Gepi / Cnrs Umr 8111
Observatoire de Paris
5, Place Jules Janssen
F-92195 Meudon Cedex
http://wwwhip.obspm.fr/gaia

Identification of simulated sources

Identification: GAIA-C2-SP-OPM-FA-049-03

Issue date: 2008-03-31

The format of the identifier of simulated sources cannot be strictly identical to the one used within DPAC, as the nature of a simulated source is known. For example, an identifier should allow to discriminate a multiple system from its components, and this, using a unique designation. We document here a more complete coding form, able to be compatible with the one adopted for the observed sources.

Prepared by:	Frédéric Arenou, Céline Reylé, Carine Babusiaux, Xavier Luri, Annie Robin, Jérôme Berthier
Verified by:	
Approved by:	CU2

Change Record

Draft 2	February 22, 2008	After comments by Céline
V1	March 11, 2008	Comments by Annie, Carine, Jérôme, Jordi. Main modif.: SSO names
V2	March 31, 2008	Comments from Philippe Redon/CNES. Modif:typos in BNF
V3	September 7, 2009	Added HII and PN, changed letters for SN and clusters

Reference Documents (RDx)

[1] C. Reylé, A. Robin, F. Arenou, and E. Grux, Universe model ICD, Technical Report GAIA-C2-SP-LAOB-CR-001-5, 2008-10-10, Laboratoire d'Astrophysique de l'Observatoire de Besançon.

[2] F. De Angeli, F. van Leeuwen, J. Hoar, and W. O'Mullane, Proposal for the object numbering scheme, Technical Report GAIA-C1-MN-IOA-FDA-002-3, 2007-03-21, Institute of Astronomy.

[3] U. Bastian, Source identifiers - Assignment and Usage throughout DPAC, Technical Report GAIA-CD-TN-ARI-BAS-020-01, 2007-06-06, ARI Heidelberg.

[4] A. A. Tokovinin, MSC - a catalogue of physical multiple stars, Astronomy and Astrophysics Supplement series, 124:75–84, 1997.

[5] W. I. Hartkopf and B. D. Mason, Addressing confusion in double star nomenclature: The Washington Multiplicity Catalog, The Environment and Evolution of Double and Multiple Stars, Proceedings of IAU Colloquium 191, Revista Mexicana de Astronomía y Astrofísica, 21:83–90, 2004.

[6] B. D. Mason, G. L. Wycoff, W. I. Hartkopf, G. G. Douglass, and C. E. Worley, The 2001 US Naval Observatory Double Star CD-ROM. I. The Washington Double Star Catalog, The Astronomical Journal, 122:3466–3471, 2001.

Acronym list

BNF Backus-Naur Form

CU Coordination Unit

DPAC Data Processing and Analysis Consortium

ICD Interface Control Document

IGSL Initial Gaia Source List

SSO Solar System Objects

WMC Washington Multiplicity Catalog

Table of contents

1 Introduction

The CU2 uses a designation of the sources both internally and externally through the Universe Model Interface Control Document (ICD) (RD 1). Initially, the identifier format was based on a region number and an object number in that region. In practice, the current (at that time) implementation coded internally that identifier as an object number plus a region number $\times 10^6$. However it then became mandatory to add the component number for multiple systems, further multiplying the number above by 10^4. At the end of cycle 2, the net result was that a long integer was no more large enough to handle all cases. Redesigning the CU2 identifier format was thus mandatory, taking into account that DPAC already proposed a source identifier format since then (RD 2; 3). The proposal described below has been implemented for the cycle 4 outputs.

2 Constraints

- The simulation needs are different from those of the data reduction. By definition, CU2 always knows whether a source is single or multiple, or is a solar system object (SSO), and it is of interest to code these characteristics in the identifier format of these objects, because CU2 has to output data concerning objects otherwise unreachable; for example, a spectroscopic binary will be one source only for DPAC while CU2 should output the data concerning the A component, the B component, and the system itself, thus needing a different identifier for the 3 sources. More generally, the other CUs may not know initially (and possibly neither at the end) the true nature of the source and may have to change the designation with time or have different designations for different observations of the same source (e.g. for SSOs). For CU2, an identifier should belong to one and one source only, for all of its observations.
- However it will be needed at some point that the CU2 identifier format is compatible with the one used for the observed DPAC data:
  - DPAC users should not have to manage two incompatible formats.
  - If data reduction uses an initial catalogue, an Initial Gaia Source List (IGSL), this will require identifiers distinct though compatible with those of the true observed data.
  - Another CU2 future need may be the following: after simulating objects on images, Gibis runs an "on-board software" which detects objects. Detected objects will not be exactly like the simulated ones (e.g. multiple systems), however beside their "TransitId", some identification similar enough to the simulated object (or at least the brightest one for multiple systems) could ease the comparison, implying the need for a compatibility between various identifier formats.
- By "compatibility " what is meant is that a CU2 identifier should be easily used in the frame of either simulated or observed data, even if they do not encode the same information.
  - For the DPAC CU$i, i \neq 2$, what is required (RD 3) is to code a region (HEALPix) and an object number in this region. For SSOs a negative value would be used; for the IGSL a special range of numbers could be used.
  - The CU2 needs to use a region (whether the scheme used is HTM, HEALPix or else does not matter here), an object number in region, the indication of multiple components, the indication of the object type (and possibly its peculiarity). As indicated above, using at least 9+4+4+4 decimal digits to code this information would produce an overflow for a 19 digits-64 bits integer.
- For the inter-CUs compatibility reason, using internally a long integer would be preferable. For practical reasons (sorting, using index in arrays) too.
- A final constraint is the following: although the identifier is used internally, it is also used externally, and there remains here and there some humans reading the produced files – who do not fully appreciate the pleasure to memorise 19-digits long integers.

3 Constraints which can be relaxed

- Nothing prevents to use an external format different from the internal coding: a long integer can be decoded and printed or read as a string. With a fast (bit logic) coding/decoding procedure, the CPU penalty will be small - actually negligible as this is done once or twice per object only, and during I/O operations much more time consuming by themselves.
- For Solar System Objects (SSO), different identifiers for the same object (corresponding to different observations) may be accepted, provided a unique internal identifier exist. As SSO have names and (not always) numbers, it looks sensible to be able to use both, with a priority on names.

4 Coding proposal

4.1 Externally

As indicated, it is much more useful to print an identifier as a meaningful string rather than as a long integer. For decades, several stellar catalogues (GSC, Tycho, etc.) have been using an identification scheme indicating an internal region number and a source number within this region. Not unexpectedly, the proposed identifier format will resemble this classical scheme.

The identifier would be a 22-char long string, padded with blanks when shorter. While examples are given Sect. 6, a more formal description using a Backus-Naur Form (BNF) is:

$$
\begin{aligned}
&< \text{identifier} > \quad ::= \quad < \text{objtype} > \; < \text{region\#} > \text{“}-\text{”} < \text{object\#} > \; [< \text{component} >] \; [< \text{variability} >] \\
&< \text{objtype} > \quad ::= \quad \texttt{“C”} \mid \texttt{“E”} \mid \texttt{“G”} \mid \texttt{“H”} \mid \texttt{“Q”} \mid \texttt{“N”} \mid \texttt{“S”} \mid \texttt{“*”} \mid \texttt{“?”} \\
&< \text{region\#} > \quad ::= \quad \texttt{“000000000..268435455”} \\
&< \text{object\#} > \quad ::= \quad \texttt{“000000..262143”} \\
&< \text{component} > \quad ::= \quad \texttt{“A..G”} \; [\texttt{“a..g”} \; [\texttt{“1..7”} \; [\texttt{“a..g”} \mid \text{“+”}] \mid \text{“+”}] \mid \text{“+”}] \mid \text{“+”} \\
&< \text{variability} > \quad ::= \quad \texttt{“V”}
\end{aligned}
$$

The object types are describedbelow while the components are described in Sect. 5. For SSO, the identifier format has to be slightly different (the objects are moving):

$$
\begin{aligned}
&< \text{identifier} > \quad ::= \quad < \text{objtype} > \; < \text{designation} > \; [< \text{component} >] \\
&< \text{objtype} > \quad ::= \quad \texttt{“a”} \mid \texttt{“c”} \mid \texttt{“p”} \mid \texttt{“s”} \\
&< \text{designation} > \quad ::= \quad < \text{name} > \mid < \text{SSO\#} > \\
&< \text{name} > \quad ::= \quad \text{“[”} < \text{SSOname} > \text{“]”} \\
&< \text{SSO\#} > \quad ::= \quad \texttt{“000000000..268435455”}
\end{aligned}
\tag{1}
$$

The object type can be a stellar system (`*`), a galaxy (`G`), a supernova (`S`), a quasar (`Q`), an exoplanet (`E`), HII region (`H`), cluster (`C`), etc.; as for the SSO: asteroid (`a`), comet (`c`), planet (`p`) or satellite (`s`). In principle, there would be no real need to code the object type or variability within the identifier. This is done because object Catalogues are frequently based on these characteristics (e.g. V*, NSV, GSH, LDN, Cl*, LEDA, MCG...): as this information about the object nature is fundamental, incorporating it in the printed identifier is thus logical [1] Besides, this allows e.g. to draw the attention, to select or to reject special object types via editor or unix commands.

4.2 Internally

The whole information above has also to be coded internally. A 64 bits-long will be used, containing the region number in the upper 32 bits and the object number in the lower 32 bits. In (RD 3), about 200 millions regions (28 bits) were indicated, with one thousand (10 bits) objects per region maximum (Baade window) at $G = 20$. Although it is unclear whether these values are optimal, we adopt them, but CU2 needs some extra margin for the object numbers (18 bits) as objects fainter than 20 and denser regions (globular clusters) have to be simulated. [2]

Besides, adapting the HTM depth may be needed. For Poisson noise reasons, the simulation of objects by the Universe Model in low density regions has to be done in large regions, but the objects may subsequently be affected to smaller regions within.

The 4 most significant bits would contain the object type, the bit sign indicating a SSO and by convention the `0000` value would not be used, allowing to check whether a given long int can be an identifier or not. As indicated above, the 28 following bits would contain the region. Then, one bit would indicate the variability, and one bit whether this is a multiple system (i.e. not a component). As described in Sect. 5, the 4 next groups of 3 bits are the 4 hierarchical levels for a multiple system. Finally the 18 least significant bits would concern the object number.

[1] The way to code star identifiers has a long history, a classical reference being the “Dictionary of Nomenclature of Celestial Objects” which is known by a user as the way to designate objects in Simbad, Vizier or ADS. For example, Vega is classically called `* 3 Lyr`, `V* alf Lyr` or `GEN# +1.00172167`, while `V* V438 Tau` is also `Cl* Melotte 22 HHJ 364` and `SV* SVS 1818`, or `[SCG94] X 63`. For SSOs, several designations occur, such as `(85) Io`, `5100 T-3` or `S/2001(107)1`. Clearly, all kind of so-called unix “special characters” may thus be found in identifiers and Simbad or other name resolvers manage this without any trouble, so this should not be considered as a potential problem if the proposed format above contain such characters. Whitespaces in a SSO name may however be considered as a problem as they may be misinterpreted as field separators in ICDs and are thus substituted by an underscore “_”.

[2] In the latter case, restricting the simulated objects to the brightest ones only will in any case be needed as more objects than that would be difficult to handle.

In theory, a simple logical AND between a CU2 identifier and `0x8FFFFFFF00040000` would quickly transform the former into a DPAC identifier. Actually, if the former uses HTM and the latter HEALPix, a further coordinates transformation would be needed.

5 Multiple systems

As indicated in (RD 4), the maximum multiplicity of a hierarchical multiple system currently found is 7 with a depth of 4 levels; this is ν Scorpii, illustrated Fig. 1. There are no special reasons to simulate more complicated systems, so we adopt a 4-levels depth and, at each level, we allow up to 7 (i.e. 3 bits are needed) sequential (i.e. non-hierarchical) components. This should be largely enough for simulation needs for which multiple systems will in most cases be reduced to double stars (or exoplanets, or multiple SSOs) with A and B components.

For the output string format, the component designation of the systems is the one used by the Washington Multiplicity Catalog (WMC) (RD 5), the root of the Washington Double Star Catalog (RD 6), which is the one adopted by the IAU. At first level, capital letters are used (from `A` to `F`), at second level lower case letters (`a..f`), at third level numbers (`1..7`).

The WMC recommends to alternate lower case letters and numbers for (currently unknown) higher levels, so we will use lower case letters at fourth level, e.g. like `Ba2b`. However, for simplicity reasons (the size of the string), we will write *systems* differently. While the WMC would write "`Ab1,Ab2`", we adopt the "`Ab+ `" writing which guarantees that the string will always be 4 characters long only and does not restrict the system to two components only.

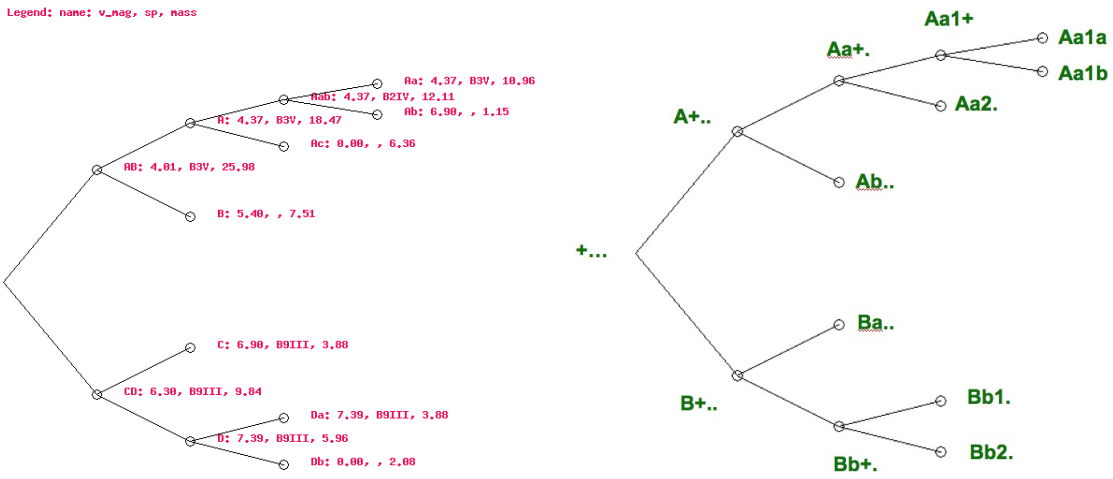

Figure 1: HR 6027 = ν Sco, probably the largest multiplicity case known so far. Left: the system as described in the MSC Catalogue (RD 4), right: the proposed designation for the nodes (systems) and leaves (stars).

6 Examples

- " `*187000548-110213` ": the star number 110213 in region 187000548
- " `*000000123-000898Ab2+V` ": the stellar system number 898 in region 123, a parent system Ab2, which is also variable (because at least one of his child is variable)
- " `G000123456-012345` ": the galaxy number 12345 in region number 123456
- " `s[Io]` ": the Jupiter satellite
- " `a000000121+` ": the system (barycentre) of the binary asteroid Hermione
- " `a[Otto_Schmidt]` ": another asteroid, by name
  " `a000002108` ": or by number