A pipeline for producing a scientifically accurate visualisation of the Milky Way

Márcia Barros[1,2], Francisco M. Couto[2], Hélder Savietto[3], Carlos Barata[1], Miguel Domingos[1], Alberto Krone-Martins[1], and André Moitinho[1]

[1] CENTRA, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[3] Fork Research, R. Cruzado Osberno, 1, 9 Esq, Lisboa, Portugal
mbarros@sim.ul.pt

Abstract. This contribution presents the design and implementation of a pipeline for producing scientifically accurate visualisations of Big Data. As a case study, we address the data archive produced by the European Space Agency's (ESA) Gaia mission. The satellite was launched in Dec 19 2013 and is repeatedly monitoring almost two billion (2000 million) astronomical sources during five years. By the end of the mission, Gaia will have produced a data archive of almost a Petabyte.
Producing visually intelligible representations of such large quantities of data meets several challenges. Situations such as cluttering, overplotting and overabundance of features must be dealt with. This requires careful choices of colourmaps, transparencies, glyph sizes and shapes, overlays, data aggregation and feature selection or extraction. In addition, best practices in information visualisation are sought. These include simplicity, avoidance of superfluous elements and the goal of producing esthetical pleasing representations.
Another type of challenge comes from the large physical size of the archive, which makes it unworkable (or even impossible) to transfer and process the data in the user's hardware. This requires systems that are deployed at the archive infrastructure and non-brute force approaches.
Because the Gaia first data release will only happen in September 2016, we present the results of visualising two related data sets: the Gaia Universe Model Snapshot (Gums), which is a simulation of Gaia-like data; the Initial Gaia Source List (IGSL) which provides initial positions of stars for the satellite's data processing algorithms.
The pipeline here presented has been approved for producing official deliverables of the ESA Gaia mission.

Keywords: Big Data, Astronomy, Visualisation

1 Introduction

To the human brain, it is easier to understand visual representations of data compared to the raw analysis of, for example, a table. Data visualisation provides an intuitive means of exploring the content of data and allows to identify patterns

of interest that might not be noticed in other way, such as blind statistical searches.

Scientific data collection has increased exponentially in the last years, and alongside with it, the challenges of managing and visualising these datasets.

Exploring and visualising large datasets has become a major research challenge[1–4]. Still, with the current knowledge and technologies it remains a challenge to visualise big datasets. One easily becomes "data stunned" and problems such as cluttering and overplotting are typical. Thus, it is crucial to follow careful approaches for visually representing the information content of these datasets.

Making such large data-sets understandable is a current concern of the scientific community and was the motivation for our work on developing a pipeline capable of creating intelligible, visually appealing and scientifically correct visualisations from large amounts of data using common hardware in an acceptable time. These challenges are particularly true for Astronomy, for which the Gaia mission is currently generating one of the largest and most complex data sets in the field.

Gaia is an ESA cornerstone mission launched on December 19th, 2013. Its main scientific goal is to understand the structure and development of our Galaxy, the Milky Way. The originality of Gaia lies on its unprecedented ability to determine stellar positions with an accuracy of a few millionths of a second of arc: the thickness of a hair at 1000 km! The mission will produce an all-sky survey of positions, distances, space motions, brightnesses and other astrophysical parameters of almost 2 billion objects brighter than magnitude 20[5, 6]. The mission is projected to have a lifetime of 5 years, with almost three now elapsed. By now, Gaia has already collected data from more than 1 billion of stars. By the end of the mission it will have produced an archive of over a 1 Petabyte. According to current definitions[7], data sets with this volume and complexity are considered big data. These are datasets which can not be managed and processed with traditional techniques within an acceptable time.

This manuscript presents our pipeline as an approach for scientific visualisation of big data, and its application to data produced from the Gaia mission of the European Space Agency (ESA). With three main modules, it receives as input tabular data, reads the data by chunks, process it, and builds a visual representation of the input data. The pipeline allows to create a visualisation for big data, with detail and quality, and because it does not load all data at once, Random Access Memory (RAM) problems are avoided. This pipeline is especially addressed to process coordinates, being useful not only to Astronomy, but also to all fields working with coordinates systems.

2 Related Work

The demand to visualise big datasets has brought to light some big data visualisation systems, for a different range of data[8–11]. However, this tools are not addressed to coordinates systems or Astronomy.

In astronomy, software such as Topcat[12], Aladdin[13] and Blender[1] are particularly used for visualising smaller datasets with up to a few million points. However, for billions of points, these tools can no longer handle the task. Paraview[14], is also used in some astronomical applications and is capable of handling large data-sets. However, its power comes from using distributed computing resources such that large data-sets will require large computer facilities.

One of the challenges in Astronomy, is to generate scientifically useful maps of our galaxy, the Milky Way, from measurements in tabular form (e.g. stellar coordinates and fluxes). The degrees of success in accomplishing this vary. In any case these maps have been built mostly using small catalogues (by today's standards) of stars, using more or less advanced techniques.

Since the dawn of photography the most common method to represent the Milky Way galaxy is through the assembly of photographic images. Barnard's Atlas of Selected Regions of the Milky Way[15], National Geographic Society Palomar Observatory Sky Survey of the northern sky[16] or Sloan Digital Sky Survey[17] are some of the examples of panoramic images of the Milky Way created from photographs. More recently, Axel Mellinger describes the construction of a colour all-sky panorama image of the Milky Way from more than 3000 individual CCD images[18].

Researchers from the Bochum University in Germany have created what is currently the photographic panorama of the Milky Way with the largest number of pixels[2].

However, there are many observables that cannot be captured directly as single images. Velocity measurements, chemical composition, variability periods, just to name a few of these, are rich in information and are usually stored in catalogues. Thus, other views of the Milky Way can be obtained by encoding these quantities in spatial maps.

By the year of 1955 in Lund Observatory, Martin and Tatjana Keskula handpainted a 2x1 meters map of the sky, with the 7000 brightest stars (Figure 1a). All of this 7000 stars were visible to the naked eye and they were painted as individual objects overlayed on a picture of the diffuse light of the Milky Way[19].

Before Gaia, ESA launched a precursor mission: Hipparcos, which was the first space mission dedicated to charting the heavens, recording the positions of stars in the sky and estimating their distances by exploring the parallax effect. It was launched in 1989 and operated until 1993. Hipparcos produced a catalogue containing the positions and distances of 118218 stars, and posteriorly the Tycho2 catalogue listing positions and motions (not distances) of over 2.5 million stars[3].

In 2002, using the information in Hipparcos and Tycho2 catalogues, a synthetic image of a full-colour all-sky star map was constructed[4]. The main goal of this map was to portray the sky as the dark-adapted naked eye perceives it.

[1] https://www.blender.org/
[2] http://aktuell.ruhr-uni-bochum.de/pm2015/pm00143.html.en
[3] http://www.cosmos.esa.int/web/hipparcos
[4] http://sci.esa.int/hipparcos/52887-the-hipparcos-all-sky map/

Image Reduction and Analysis Facility (IRAF) software was used to generate the star map. A full-colour photography of the constellation of Orion was used as benchmark. The final result was an image of the sky with a size of 20000×10000 pixels (Fig1b).

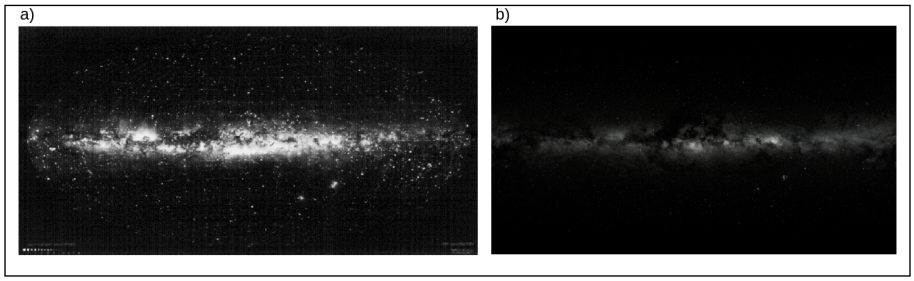

Fig. 1. a) hand-painted map of the Milky Way; b) synthetic full-colour all-sky star map from Hipparcos data

The two cases given above illustrate the possibilities and limitations of technical approaches for producing maps of the Milky Way. The Lund map[19], was based on 7000 stars and could be painted by hand. The Hipparcos map, with 2.5 million stars required a common computer and software used by astronomers. This last approach does not scale to the Gaia products, due to the amount of collected data (more than one billion of stars).

Here we present a first prototype of a pipeline for serious information visualisation, following best practices with include setting adequate colour-maps, simplicity and avoidance of superfluous elements.

3 Data description

The basic type of data required by our pipeline as input are sets of coordinates, a latitude and a longitude which mark the position of objects (for example, the position of all cars, the position of all zoos, the position of all stars in the Milky Way). We used the coordinates of stars in the Milky Way.

Gaia is a system of two telescopes with three instruments delivering different kinds of data: The Astrometric Instrument delivers positions that are used for computing transverse motions and distances; The Radial Velocity Spectrometer delivers high resolution spectra that allow computing the velocity of stars along the line-of-sight though doppler shift measurements; The Photometric Instrument delivers low-resolution spectra in the red and blue range, which are used for computing the apparent luminous fluxes (magnitudes), temperature, mass and chemical composition of stars. Gaia detects astronomical sources on-board and transmits only small rasters around each source to Ground stations. Process-

ing this telemetry into the final science-ready catalog and intermediate products is then done by the Gaia Data Processing and Analysis Consortium (DPAC).

The data used in the pipeline here described consists of the science ready catalog of stellar positions and physical parameters produced by the DPAC[5].

Because the first Gaia data release will be in September 2016,to the development of our pipeline, we used other data sources: Gaia Universe Model Snapshot (GUMS) and the Initial Gaia Source List (IGSL). These are the standard test data for DPAC development.

GUMS is a computational simulation of the objects and characteristics that Gaia will potentially observe. It has approximately 2 billions simulated objects, including longitude and latitude, magnitudes and other parameters[20].

IGSL is the starting point for object positioning in the Gaia data processing. It was compiled from large scale public catalogs of stellar positions and a few specific catalogs provided by the DPAC. It has over 1 billion entries, corresponding to a file with approximately 250 Gigabytes, with most of parameters that Gaia will publish[21].

4 Approach

In Astronomy, as in other sciences, Python has become a widely used programming and analysis language. In the case of astronomy there are a huge number of packages covering from basic astronomical functions such as coordinate transformations to advanced modeling and statistical analysis.

The pipeline represented in Figure 2 has been designed with the Big Data challenge in mind. The input data is tabular and contains at least a pair of coordinates per source followed by the source's attributes. Module 1 reads the data by chunks of, for example, 10^4 lines at a time. Module 2 processes each chunk and does the necessary computations (described below) for producing the 2D matrix of pixel positions and values that will be visualized. Finally, module 3 renders the visualization.

We now describe the pipeline in more detail. Since the pipeline will be openly available and likely to be used at least within the Gaia community (now with approximately 600 researchers), we have opted to also develop it in Python.

Module 1 reads the data in chunks for avoiding overloading by too many data. This was done using pandas, an open source library which provides data structures and data analysis tools for Python. Each chunk is passed to Module 2.

Module 2 processes the lines in each chunk, preparing the data for being rendered into an image. In Astronomical data, coordinates can be represented in different spherical systems. The most common systems are centered on the Sun but have different orientations of their poles and equators. These include the Equatorial system (equator of the Earth projected on the celestial sphere) the Galactic system (equator aligned with the plane of the Milky Way) or Ecliptic

[5] http://www.cosmos.esa.int/web/gaia/release

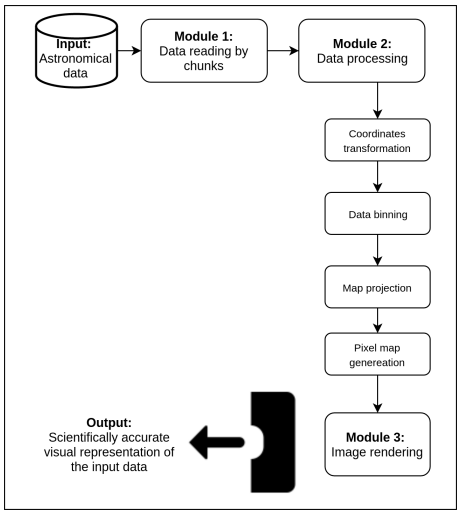

Fig. 2. Data processing pipeline for the astronomical case study. Astronomical tabular data as input. Module 1: data reading by chunks. Module 2: process the data, performing coordinates transformation, data binning, map projection application and a pixel map generation. Module 3: uses Python image techniques to draw the pixels map generated by Module 2. The output of this pipeline is a scientifically accurate image of the input data.

(equator aligned with the plane of the solar system) systems. Users will often require producing maps from any of these different perspectives. Catalogs will usually only provide coordinates in one system. Thus, a first step is to perform an optional transformation of coordinates.

The Gaia catalogue will list equatorial coordinates which produce a twisted view of the plane of the Milky Way (see Figure 3). Thus we have applied a equatorial to galactic coordinate transformation (step 1 of module 2). Astronomical coordinate transformations, such as the one applied are provided by the Python Astropy package.

Astropy provides a large panoply of astronomy-related functionalities. It supports different file formats such as flexible image transport system (FITS) files, Virtual Observatory (VO) tables, and ASCII table. It allows unit and physical quantity conversions, physical constants specific to astronomy, celestial coordinate and time transformations, world coordinate system (WCS) support, generalized containers for representing gridded as well as tabular data, and a framework for modelling and statistical analysis[22].

Step 2 of module 2 does a partitioning of the coordinate space for aggregating data and computing statistics to be visualized. The simplest case would be a 2D source density map. While aggregating data in a cartesian space can be easily done (among other possibilities) by dividing space into equal area squares (or hyper-cubes for higher dimensions), dividing a sphere in equal area sections is

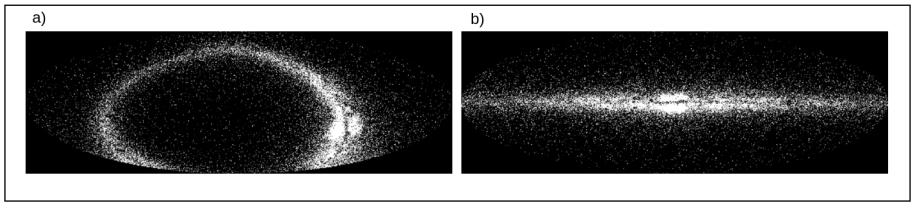

Fig. 3. Sample of GUMS dataset represented in a) equatorial coordinates; and b) galactic coordinates.

not trivial. One way to accomplish this, popular in Astronomy, is to divide the sphere using Hierarchical Equal Area isoLatitude Pixelation (HEALPix)[23].

HEALPix is a mathematical structure which supports a suitable discretisation of functions on a sphere at sufficiently high resolution, and provides an index for easy, fast and accurate statistical astrophysical analysis of massive full-sky data sets. It has three major characteristics: the sphere is hierarchically tessellated into curvilinear quadrilaterals, the area of each pixel at a given resolution is identical, and pixels are distributed on lines of constant latitude. Using HEALPix allows to our pipeline to distribute each star in the correct position of a matrix, and likewise to produce correct statistical samples.

HEALPix software is available in several programming languages such as C, C++, Fortran, JAVA and Python. We used Healpy, the python package for manipulating healpix maps. Healpy allows to chose the number of healpixels (nside) in which the coordinates will be distributed. This number has to be a power of 2, limited to a maximum of 2^{23}. A nside of x creates $12 * (x^2)$ healpixels. Thus, through healpy, the galactic coordinates are transformed into healpixels numbers, i.e., each pair longitude-latitude is replaced by a number representing the correspondent healpixel for that coordinate, for a given nside. A reverse transformation from healpixs to angular coordinates gives the central coordinates of each healpix with correctly computed statistics associated to those coordinates.

The third step in module 2 is to project on a flat cartesian plane the map previously computed on a spherical representation. A sphere can not be represented on a plane without distortion[24], thus, one must analyze the trade-offs between different coordinate projection schemes. The Hammer projection, adopted in this work, is known for being an equal-area projection that reduces distortions towards the edges of the map. The projection results in x,y positions in a 2:1 ratio with x confined to (-1,1).

The last step of module 2 is to scale the x,y coordinates to a matrix with a given range and pixel size, and assign each x,y pair to a pixel. The size of our matrix is the desired size of our image. The number stored in each position of the matrix correspond to the statistics of stars agglomerated in that position.

The third module of this pipeline uses the matrix created by module 2 for rendering the final image. The Python Imaging Library (PIL)[25] was used. It provides modules that can perform many image manipulation and processing

operations in Python and draws images pixel by pixel. Gizeh, a library for vector graphics based on PyCairo, was also used for drawing shapes, such as circle with transparencies for representing brighter and closer stars.

The images presented in this work follow from the initial tests of the pipeline and represent the number density of sources within each Healpix. Other images are being developed, such as maps of integrated stellar fluxes as well as kinematic and chemical maps of the Milky Way.

For the density, several colour schemes have been tested (see Figure 4). We find that in general, grayscale maps provide the best representations, consistent with photographic plates and free of misleading artifacts that rainbow like colourmaps are known to produce [26]. White was attributed to the pixels with highest densities, and back to the pixels with lowest densities.

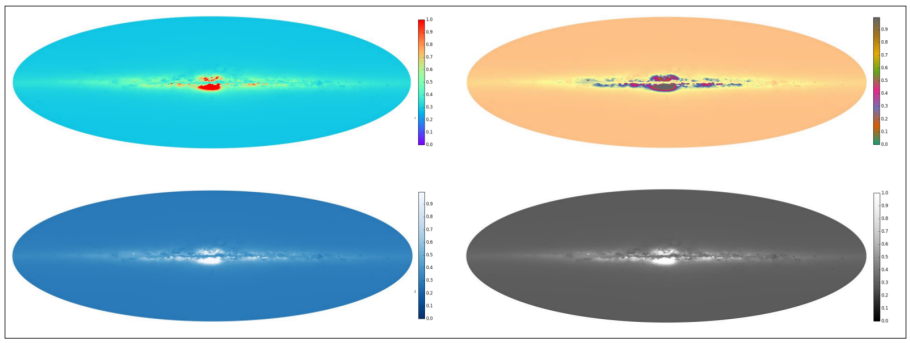

Fig. 4. GUMS data with different colormaps, from matplotlib python library. Left top - rainbow; Right top - Accent; Left bottom - Blues_r; Right bottom - Greys_r.

5 Results

The main goal of this work was to create a realistic map of the Milky Way, with Gaia data, inspired by the Hipparcos map (Figure 1a). However, unlike Hipparcos, with considerably less stars (approximately 118000), the Gaia catalog will list almost 2 billion stars. As explained in Section 3, here we test the pipeline using the GUMS and IGSL datasets. Our pipeline is now able to process one coordinate in approximately 1.5×10^{-5} seconds.

Figure 5a represents approximately 2 billion of stars from Gaia Universe Model Snapshot (GUMS) simulation, processed by our pipeline. It is a density map of the Milky Way, in galactic coordinates, with a size of 6000×2000 pixels, with the densities computed in HEALPix, and displayed in a Hammer projection. The whitest zones correspond to the larger densities and the darkest zones to lower densities. This colour map clearly displays the known features of the simulated Milky Way such as the higher concentration in the center and the

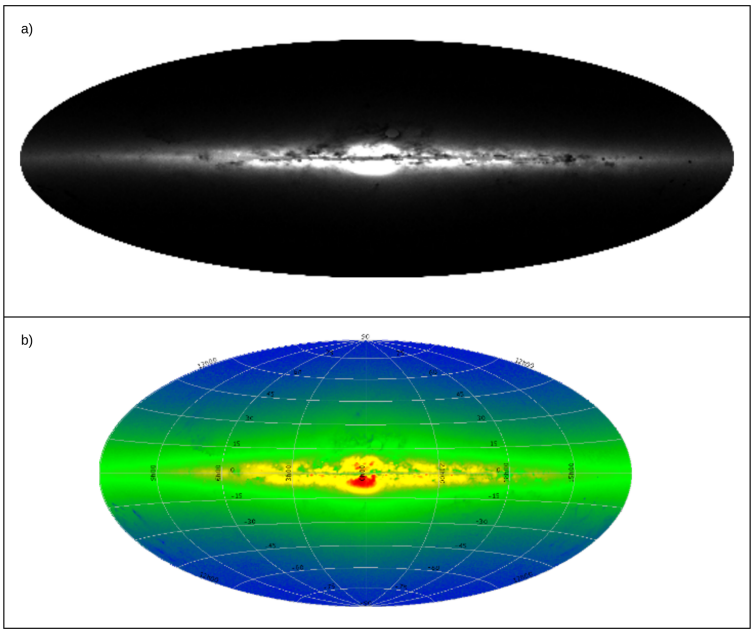

Fig. 5. GUMS dataset visualisations created by a) our pipeline; and b) another visualisation system.

darker band in the galactic equator representing the effects of light absorption by interstellar matter.

Comparing our image (Figure 5a) with another visualization of the same data (Figure 5b)[6][20], the first one, resembles a realistic photograph of Milky Way, instead figure 5b, although informative, in not realistic and introduces the false impression of a delimited high density region coded in red.

The second dataset used for tests was the Initial Gaia Source List (IGSL) and is based on real sky measurements. The result is shown in Figure 6a. This image has the same technical characteristics of Figure 5a (density map, resolution, gray scale colourmap). Here, other structures can be identified such as the Magellanic Clouds (Figure 6b, left) and small clusters of stars (Figure 6b, right). Figure 6c shows an alternative visualization of the same data[7][21]. Both figures, 6a and 6c, reveal the same circular patterns. These patterns are footprints of the ground based imaging surveys from which the IGSL was built. Again, it is seen how the choice of a colourmap can introduce the false impression of delimited high density regions (coded in red). Figure 6a, as Figure 5a, transmits the feeling of seeing the actually real night sky, this being one of the goals of this project: to create an image informative, real, appealing to our sight.

[6] http://gaia.esac.esa.int/tap-server/StatGraph?TABLE=public.g10_mw&TYPE=DENSITY
[7] http://gaia.esac.esa.int/tap-server/StatGraph?TABLE=public.igsl_source&TYPE=DENSITY

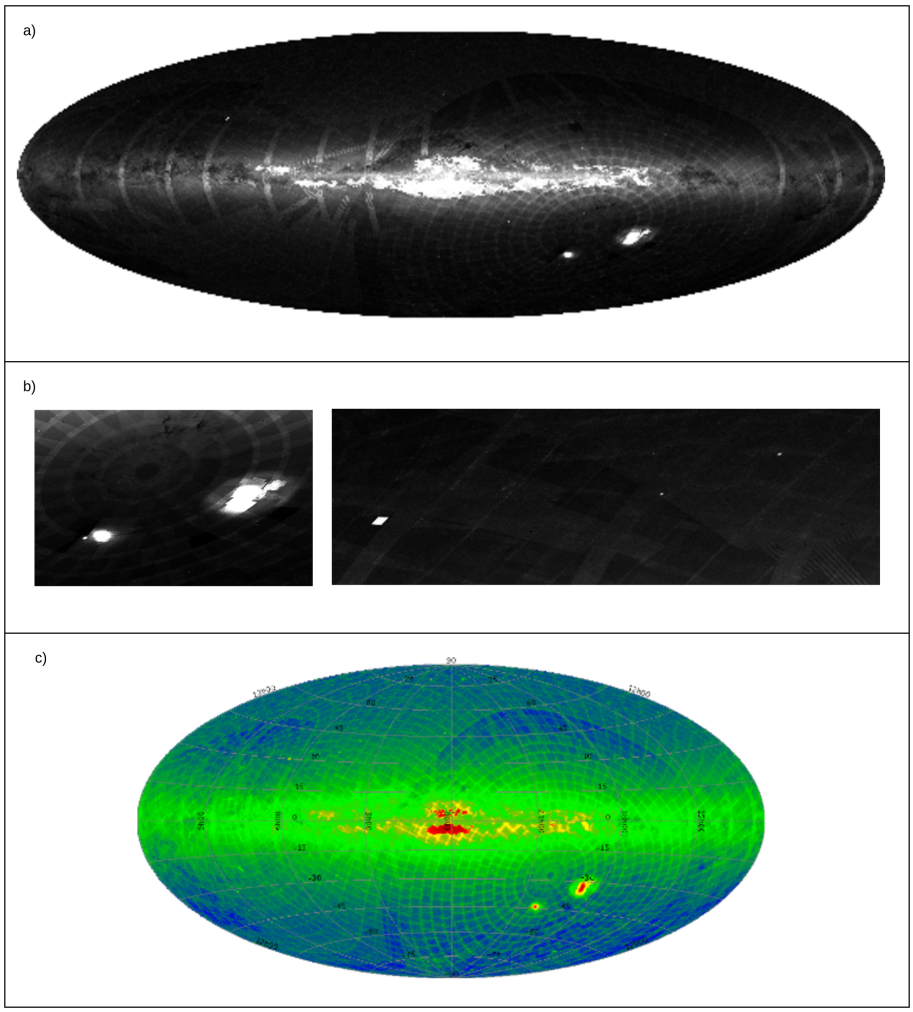

Fig. 6. IGSL dataset visualisations created by a) our pipeline; b) details of a), with Magellanic Clouds on the left and stars agglomerations on the right ; and c) another quick look visualisation system.

6 Conclusion

Along this paper we have discussed some challenges of big data visualization. Our work contributed to the development of a new pipeline that we demonstrated to be feasible for visualizing Big scientific Data. In the astronomical case study we presented, the pipeline showed to be adequate for reading, processing and visualizing in a scientifically accurate form, the massive amount of data produced by Gaia mission.

The pipeline can already deal with great amounts of data (billions of points) stored in hard drive with no need of bulk loading to RAM. However there are issues to solve. For example, it takes almost five hours to read and process all the data. We intend to reduce this value with threading and multiprocessing methods.

For the future we intend to create images using other parameters such as stellar magnitude, colour, astrometric, kinematic and chemical indicators. Another development, is to work on overrepresentation analysis to find which features are discriminatory of a set of points (selected by the user) with respect to the whole dataset. The goal is to help the user understand in real time the shared semantics behind a given selection of points, for example by using a Bayesian approach[27].

7 Acknowledgments

This work was supported by the European Union's Seventh Framework Programme (FP7-SPACE-2013-1) under grant agreement n. 606740 (GENIUS). Work partially supported by FCT funding of the CENTRA (UID/FIS/00099/2013) and LaSIGE (UID/CEC/00408/2013).

References

1. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: A survey of the state of the art. arXiv preprint arXiv:1601.08059 (2016)
2. Heer, J., Kandel, S.: Interactive analysis of big data. XRDS: Crossroads, The ACM Magazine for Students **19**(1) (2012) 50–54
3. Godfrey, P., Gryz, J., Lasek, P.: Interactive visualization of large data sets. Technical report, Technical Report EECS-2015-03 March 31 2015. Department of Electrical Engineering and Computer Science. York University. Toronto, Ontario. Canada (2015)
4. Shneiderman, B.: Extreme visualization: squeezing a billion records into a million pixels. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM (2008) 3–12
5. Lindegren, L., Babusiaux, C., Bailer-Jones, C., Bastian, U., Brown, A.G., Cropper, M., Høg, E., Jordi, C., Katz, D., Van Leeuwen, F., et al.: The gaia mission: science, organization and present status. Proceedings of the International Astronomical Union **3**(S248) (2007) 217–223
6. Eyer, L., Holl, B., Pourbaix, D., Mowlavi, N., Siopis, C., Barblan, F., Evans, D., North, P.: The gaia mission. arXiv preprint arXiv:1303.0303 (2013)
7. Snijders, C., Matzat, U., Reips, U.D.: Big data: Big gaps of knowledge in the field of internet science. International Journal of Internet Science **7**(1) (2012) 1–5
8. Quan, D., Karger, R.: How to make a semantic web browser. In: Proceedings of the 13th international conference on World Wide Web, ACM (2004) 255–265
9. Rutledge, L., Van Ossenbruggen, J., Hardman, L.: Making rdf presentable: integrated global and local semantic web browsing. In: Proceedings of the 14th international conference on World Wide Web, ACM (2005) 199–206

10. Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., Sheets, D.: Tabulator: Exploring and analyzing linked data on the semantic web. In: Proceedings of the 3rd international semantic web user interaction workshop. Volume 2006., Athens, Georgia (2006)
11. Liu, Z., Jiang, B., Heer, J.: immens: Real-time visual querying of big data. In: Computer Graphics Forum. Volume 32., Wiley Online Library (2013) 421–430
12. Taylor, M.B.: TOPCAT & STIL: Starlink Table/VOTable Processing Software. In Shopbell, P., Britton, M., Ebert, R., eds.: Astronomical Data Analysis Software and Systems XIV. Volume 347 of Astronomical Society of the Pacific Conference Series. (December 2005) 29
13. Stafford, J.A., Richardson, D.J., Wolf, A.L.: Aladdin: A tool for architecture-level dependence analysis of software systems. Technical report, DTIC Document (1998)
14. Squillacote, A.H., Ahrens, J.: The paraview guide. Volume 366. Kitware (2007)
15. Barnard, E.E., Frost, E.B., Calvert, M.R.: A Photographic Atlas of Selected Regions of the Milky Way. Washington, DC: Carnegie Institution (1927)
16. Minkowski, R., Abell, G.: The national geographic society-palomar observatory sky survey. Basic Astronomical Data: Stars and Stellar Systems **1** (1963) 481
17. York, D.G., Adelman, J., Anderson Jr, J.E., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al.: The sloan digital sky survey: Technical summary. The Astronomical Journal **120**(3) (2000) 1579
18. Mellinger, A.: A color all-sky panorama image of the milky way. Publications of the Astronomical Society of the Pacific **121**(885) (2009) 1180
19. Lundmark, K.: Copernicus and luther: A critical study. yearbook of the Lund Astronomical Society Astronomiska Sallskapet Tycho Brahe (1955) 87–93
20. Robin, A., Luri, X., Reylé, C., Isasi, Y., Grux, E., Blanco-Cuaresma, S., Arenou, F., Babusiaux, C., Belcheva, M., Drimmel, R., et al.: Gaia universe model snapshot-a statistical analysis of the expected contents of the gaia catalogue. Astronomy & Astrophysics **543** (2012) A100
21. Smart, R., Nicastro, L.: The initial gaia source list. Astronomy & Astrophysics **570** (2014) A87
22. Robitaille, T.P., Tollerud, E.J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A.M., Kerzendorf, W.E., et al.: Astropy: A community python package for astronomy. Astronomy & Astrophysics **558** (2013) A33
23. Gorski, K.M., Hivon, E., Banday, A., Wandelt, B.D., Hansen, F.K., Reinecke, M., Bartelmann, M.: Healpix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. The Astrophysical Journal **622**(2) (2005) 759
24. Rassias, G.M.: The mathematical heritage of CF Gauss. World Scientific (1991)
25. Vaingast, S.: Beginning Python visualization: crafting visual transformation scripts. Apress (2014)
26. Borland, D., Taylor II, R.M.: Rainbow color map (still) considered harmful. IEEE computer graphics and applications (2) (2007) 14–17
27. Machado, C.M., Freitas, A.T., Couto, F.M.: Enrichment analysis applied to disease prognosis. J. Biomedical Semantics **4** (2013) 21