Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

Objective ICT-2009.4.1 b) – “Advanced preservation scenarios”

D5.3v1: Propagation of interdependent quantities in the calculation of luminosities of galaxies

Deliverable Co-ordinator: José Enrique Ruiz (IAA)

Deliverable Co-ordinating Institution: IAA

Other Authors: Dr. Lourdes Verdes-Montenegro (IAA); Susana Sánchez (IAA)

Document Identifier:	Wf4ever/2010/D5.3v1/v1.0	Date due:	23/11/2011
Class Deliverable:	Wf4ever 270192	Submission date:	**30/11/2011**
Project start date:	December 1, 2010	Version:	V1.0
Project duration:	3 years	State:	Final
		Distribution:	Public

Wf4Ever Consortium

This document is a part of the Wf4Ever research project funded by the IST Programme of the Commission of the European Communities by the grant number FP7-ICT-2007-6 270192. The following partners are involved in the project:

Intelligent Software Components S.A. Edificio Testa Avda. del Partenón 16-18, 1º, 7ª Campo de las Naciones, 28042 Madrid Spain Contact person: Dr. Jose Manuel Gómez-Pérez E-mail address: jmgomez@isoco.com	**University of Manchester** Department of Computer Science, University of Manchester, Oxford Road Manchester, M13 9PL United Kingdom Contact person: Professor Carole Goble E-mail address: carole.goble@manchester.ac.uk
Universidad Politécnica de Madrid Departamento de Inteligencia Artificial Facultad de Informática, UPM 28660 Boadilla del Monte, Madrid Spain Contact person: Dr. Oscar Corcho E-mail address: ocorcho@fi.upm.es	**University of Oxford** Department of Zoology University of Oxford South Parks Road, Oxford OX1 3PS United Kingdom Contact person: Dr. Jun Zhao / Professor David De Roure E-mail address: jun.zhao@zoo.ox.ac.uk, david.deroure@oerc.ox.ac.uk
Poznań Supercomputing and Networking Center Network Services Department Poznań Supercomputing and Networking Center Z. Noskowskiego 12/14, 61-704 Poznan Poland Contact person: Dr.Raúl Palma de León E-mail address: rpalma@man.poznan.pl	**Instituto de Astrófísica de Andalucía** Dpto. Astronomía Extragaláctica Instituto Astrofísica Andalucía Glorieta de la Astronomía s/n 18008 Granada, Spain Contact person: Dr. Lourdes Verdes-Montenegro E-mail address: lourdes@iaa.es
Leiden University Medical Centre Department of Human Genetics Leiden University Medical Centre Albinusdreef 2, 2333 ZA Leiden The Netherlands Contact person: Dr. Marco Roos E-mail address: M.Roos1@uva.nl	

Change Log

Version	Date	Amended by	Changes
0.1	10-11-2011	José Enrique Ruiz	Initial Draft
0.2	21-11-2011	José Enrique Ruiz	Writing for internal group amendments
0.3	22-11-2011	Susana Sánchez	Corrections on the text Executive summary and Conclusions added
0.4	23-11-2011	Lourdes Verdes-Montenegro	Corrections on the text
0.5	24-11-2011	Susana Sánchez José Enrique Ruiz	Minor corrections on the text
1.0	01-12-2011	José Enrique Ruiz	Minor corrections based on QA Final state

Executive Summary

This document describes the development of the first Golden Exemplar proposed in Work Package 5: Workflow Astronomy Preservation. It pertains to the deployment of a Research Object (RO) associated to the study of the propagation of astrophysical quantities in the calculation of luminosities of galaxies. The main goal of this deliverable D5.3v1 is to produce the workflows and RO for the first Golden Exemplar. These may be accessed publicly in the MyExperiment portal, where the whole RO has been uploaded as MyExperiment Pack[1]. Preservation aspects have been preliminarily explored now while they will be explored in more detail in future versions D5.3v2/v3.

The scientific experiment represented by the RO produced is related with the curation of a local user database storing values of physical properties needed in the multi-wavelength study of a sample of the most isolated galaxies in the local universe. The content of this database relies on a set of basic experimental values retrieved from external data repositories and combined by means of mathematical equations. The update of the external data repositories has an impact on the preservation of the digital experiment, since it is essential to know how and when those repositories are updated, so that the propagation of the changes through the existing internal relation among the data can be triggered and registered.

We provide a detailed description of the Golden Exemplar including information about the implementation process of the RO and the scientific results obtained after the enactment of the workflows. Further, we provide a discussion about how the use of the RO will have an impact on the present curation tasks undertaken in the group, preservation and versioning issues of the RO as well as requirements for annotations applied to the RO. Finally we provide several scenarios of use, one focused in the creation and publication of the RO (submission), other focused in discovery and re-use of the RO (dissemination), and a brief discussion related to the structure and organization of the RO content storage (archival).

[1] http://www.myexperiment.org/packs/231.html

Table of contents

Executive Summary 4

1. Introduction 7

2. Materials and methods 8

2.1 The AMIGA catalogue 8

2.2 The HyperLEDA database 8

2.3 The calculation of the luminosity 8

2.4 Strategy and workflow building 9

2.5 Research Object Management 13

3. Results 14

3.1 Comparison of retrieved and final calculated physical properties 14

3.2 Propagation of interdependent physical properties 14

4. Discussion 15

4.1 Impact 15

4.2 Preservation and versioning 16

4.3 Annotations 16

4.4 Scenario for submission, dissemination and archival 17

4.4.1 Submission 17

4.4.2 Dissemination 18

4.4.3 Archival 18

5. Conclusions 19

6. References 20

Appendix A – Graphical representation of nested workflows 21

Appendix B – Research Object structure and content 24

Appendix C – RO Management Tool session commands 25

Appendix D – Manifest file for Research Object 26

Appendix E – Plots of workflows results 27

List of Figures

Figure 1: Diagram followed in the process of calculating luminosities and updating local catalogue 10
Figure 2: Workflow for gathering physical quantities from the HyperLEDA 11
Figure 3: Workflow for propagation of physical quantities provided by the user 11
Figure 4: Workflow for calculation of luminosities with quantities coming from HyperLEDA 12
Figure 5: Running workflow for calculation of luminosities with quantities coming from HyperLEDA 12
Figure 6: Workflow for comparison of physical quantities and update of user catalogue 13
Figure 7: Progress report for the workflow of propagation of quantities obtained from HyperLEDA 15
Figure 8: Workflow for parsing the GalaxyNames file 21
Figure 9: Workflow for extracting property values from HyperLEDA 21
Figure 10: Workflow for extracting J2000 coordinates from HyperLEDA 22
Figure 11: Workflow for calculating distances of galaxies 22
Figure 12: Workflow for calculating the correction for total B magnitude (BT) 23
Figure 13: Workflow for calculating luminosities of galaxies 23
Figure 14: Differences in the galactic extinction (Ag) in the comparison process 27
Figure 15: Differences in the total B magnitude (BT) in the comparison process 27
Figure 16: Differences in the log of axis ratio (logr25) in the comparison process 28
Figure 17: Differences in the radial velocities in the comparison process 28
Figure 18: Differences in the uncertainty of the radial velocities in the comparison process 29
Figure 19: Differences in the total corrected B magnitude (mB-corr) in the comparison process 29
Figure 20: Differences in the distances in the comparison process 30
Figure 21: Differences in the luminosities in the comparison process 30
Figure 22: Differences in the luminosities when only affected by galactic extinction (Ag) 31
Figure 23: Differences in the luminosities when only affected by total B magnitude (BT) 31
Figure 24: Differences in the luminosities when only affected by the log of axis ratio (logr25) 32
Figure 25: Differences in the luminosities when only affected by the velocities 32
Figure 26: Differences in the morphological types used 33
Figure 27: Differences in the luminosities when only affected by morphological types 33
Figure 28: Differences in the luminosities w.r.t. differences in other quantities 34

1. Introduction

The AMIGA[2] project is an international scientific collaboration led from the Instituto de Astrofísica de Andalucía (CSIC). It focuses on a multi-wavelength analysis of the interstellar medium of an statistically significant sample of isolated galaxies, in order to provide a pattern of behaviour to the study of galaxies in denser environments. Tabular data representing specific properties of the AMIGA galaxies are registered and curated by the AMIGA group, published in astronomical journals and provided to the community via Virtual Observatory (VO) web services and a web query interface. Many of these data are calculated through mathematical equations involving as input basic experimental values coming from widely known catalogues of galaxies. The values provided by these catalogues are subject to (sometimes periodic) update, which requires a continuous vigilance for curation tasks in the AMIGA group. It is essential to know how different are the releases and when they are updated, so that the propagation of the changes in the input data values through the existing internal relations among the data can be triggered and registered.

This is a suitable case to be studied in the context of the Wf4Ever project. It is deeply related with versioning and preservation issues concerning local and external resources that are beyond the control of the user. Most of the tasks involved in the curation process may be automated in a workflow, enabling reproducibility, re-use and re-purpose for similar cross-boundary use cases dealing with up to date and controlled information. Moreover, the propagation of data values through complex calculations needs registration of provenance in order to evaluate integrity and authenticity in the whole process.

In the case described in this document, the corrected values of apparent magnitudes (brightness) and distances of an ensemble of galaxies are calculated using basic experimental values provided by the HyperLEDA[3] database. From these two calculated quantities, values of intrinsic luminosities are derived for each galaxy based on a mathematical dependence, so any modification in the values where the brightness or distances rely on will affect the final luminosities. We have developed several workflows that interrogate and gather values of physical properties from the HyperLEDA catalogue for all the AMIGA galaxies, calculate the corrected apparent magnitudes, distances and intrinsic luminosities of the galaxies, compare values stored in a local database coming from previous releases of HyperLEDA with those calculated in the actual process, and register the new values keeping a version of the old ones. The execution of the workflows is done in the local desktop of the user with Internet access to the external data sources and a python-scripting running environment.

A description of the methodology, resources and Wf4Ever tools is provided in Sect. 2 and results are presented in Sect. 3. A discussion about how this development impacts on present curation techniques and research on the AMIGA group, RO management as well as preservation issues can be found in Sect. 4 together with a scenario for submission, archival, and dissemination situations that could be useful to analyse the Wf4Ever RO model under different perspectives. Sect. 5 is dedicated to conclusions.

[2] http://amiga.iaa.es

[3] http://leda.univ-lyon1.fr

2. Materials and methods

2.1 The AMIGA catalogue

The AMIGA project is performing a multi-wavelength study for a sample of the most isolated galaxies in the local universe. The intrinsic luminosity of 1051 galaxies in the Johnson B-band is calculated from basic experimental data through complex mathematical equations. All the data involved in the process of the calculation of the luminosities are stored in a local relational MySQL database, available for all the members of the AMIGA group. The luminosities as well as other physical properties including optical, IR, radio line and continuum measures are published in astronomical journals and provided to the community via VO services and a web based query interface [5] Much of the scientific research produced in the AMIGA group relies on the curation of all the data involved in the calculation of the final physical properties provided to the community, the last release dated on June 2010.

2.2 The HyperLEDA database

The HyperLEDA (Hyper Linked Extragalactic Databases and Archives) database [6] provides most of the data needed for the calculation of luminosities for all the galaxies in the AMIGA catalogue. At present the database contains over 3 million objects, out of them 1.5 million galaxies with a high level of confidence. HyperLEDA offers a web based query interface that, given the name or coordinates of a galaxy, allows the retrieval of homogenized astrophysical data related with the physics and evolution of galaxies [3] The parameters and data values involved in the calculation of luminosities for the AMIGA galaxies are the following: the heliocentric velocity (V_{hel}), the galactic dust extinction (A_g), the axis ratio of the isophote 25 mag/arcsec^2 in the B-band, the apparent total B magnitude (B_T) and the morphological type (*t*).

2.3 The calculation of the luminosity

The optical luminosity in B-band L_B for a galaxy can be expressed in function of solar luminosities $L_\odot$ applying the following equation.

$$log\left({}^{L_B}\!/_{L_\odot}\right) = 11.95 + 2\, log[D(Mpc)] - 0.4 m_{B\text{-}corr}$$

As it can be seen, there are two physical parameters involved in the calculation: the distance of the galaxy in units of Mega parsecs *D (Mpc)*, and the corrected apparent B magnitude $m_{B\text{-}corr}$. These, in turn may be calculated applying the following equations.

$$D\,(Mpc) = V_{vir}/H_0$$

$$m_{B\text{-}corr}{=}B_T\text{-}A_g\text{-}A_i\text{-}A_K(t)V_{hel}/10000$$

V_{vir} is the radial velocity of the galaxy measured in the reference system of the Virgo cluster and H_0 is the Hubble constant, where a value of 75 km/s Mpc^{-1} has been used. B_T is the apparent total B magnitude, A_g is the correction for the dust extinction produced by our own galaxy, A_i is the internal extinction value and A_K is the K correction coefficient calculated as a function of the morphological type *t.* The radial velocity of the galaxy measured in the solar reference system is expressed as V_{hel}.

Values for B_T, A_g and V_{hel} are provided by the HyperLEDA database, while the values for the internal dust extinction A_i, V_{vir} radial velocity and A_K correction need to be estimated through more complex calculations. In particular the A_i correction coefficient is a function of the morphological type of the galaxy, the value for the axis ratio of the isophote 25 mag/arcsec^2 in the B-band (*logr25*) and a fixed coefficient *C* accounting for second order effects. The V_{vir} radial velocity needed in the calculation of the distance is a function of V_{hel}, the coordinates of the galaxy and the coordinates of the centre of the Virgo cluster. The A_K corrections are just a collection of fixed values defined for every morphological type. Values for the *logr25* are provided by the HyperLEDA database while morphological types are provided by the AMIGA group, since they are continuously revised through visual inspection of images.

2.4 Strategy and workflow building

In order to alleviate the complexity in the curation process of the luminosities of the AMIGA galaxies and improve the degree of reproducibility, we have decided to follow the approach exposed in Figure 1 for the maximum automation of the whole process. We have developed several workflows that are represented in different colours in the diagram (yellow, green and blue) while the data provided by the user are represented in orange. In a first step, a workflow for gathering the physical properties provided by HyperLEDA database can be executed providing a list of galaxy names. At this moment, a second workflow for comparing the extracted properties with those existing in the user database can be used to determine potential differences. The user then chooses whether to continue or not with the process, in that case a third workflow is executed for calculating the distances, corrections and luminosities of galaxies based on the obtained new values. Finally, the previously used workflow for comparison previous and new values can be applied this time to compare the final calculated properties, as the distances, corrected apparent B magnitude $m_{B\text{-}corr}$ and luminosities. Inspection of the results of this workflow is needed in order to decide whether a new release for all the quantities involved should be registered as curated and hence the old one archived.

It is interesting to note the separation of the whole process in several workflows, mainly due to the fact that at some moments the user needs to make a decision based on the inspection of the differences found between local stored values and values provided by the HyperLEDA database or issued from calculations. The decisions are actually made when differences are greater that certain *thresholds* which may vary from user to user. It is then important to provide a *dry run* for the update of local stored values and not to automate this task, since it should be executed based only on the user's discretion.

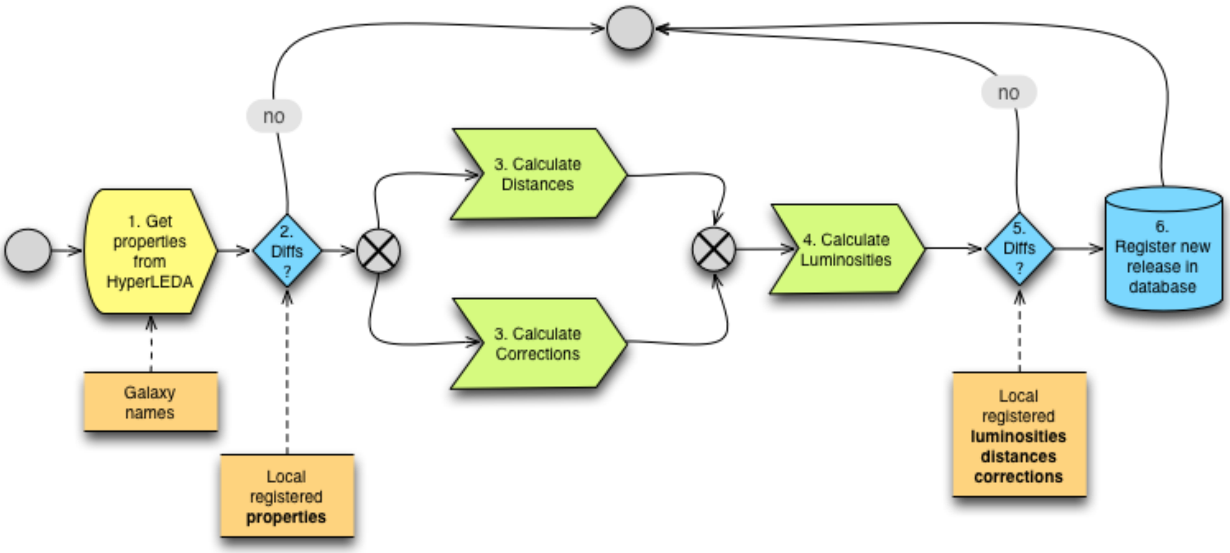

no
no
1. Get properties from HyperLEDA
2. Diffs ?
3. Calculate Distances
3. Calculate Corrections
4. Calculate Luminosities
5. Diffs ?
6. Register new release in database
Galaxy names
Local registered **properties**
Local registered **luminosities distances corrections**

Figure 1: Diagram followed in the process of calculating luminosities and updating local catalogue

The workflows have been developed using the Taverna desktop workbench version 2.3 provided by the open source suite Taverna Workflow Management System[4]. The workflow for gathering physical quantities from the HyperLEDA database is represented in Figure 2, where a more detailed representation of the comprised nested workflows can be found in Appendix A. The user provides an ASCII text file with a list of galaxy names, and the workflow queries the HyperLEDA database and parses the response with the help of regular expressions included in simple Python scripts. The result of the workflow is a series of ASCII text files with the values of the equatorial coordinates in J2000 epoch, the velocities in km/s, the dust extinction coefficient A_g, the axis ratio of the isophote 25 mag/arcsec^2 *logr25* and the apparent total B magnitude B_T. These values may be used as input data in the workflow represented in Figure 3, and a more detailed representation of the comprised nested workflows can be found in Appendix A. This second workflow propagates those input values in the process of calculation of the final physical properties with the help of Python scripts. The result of this workflow is a series of ASCII text files containing the distances, corrected apparent B magnitude $m_{B\text{-}corr}$ and luminosities for each of the galaxies present in the initial file of galaxy names. These two workflows can be merged into a single bigger one that accounts for the previous entire process described and which is represented in Figure 4 and Figure 5. A more detailed representation of the comprised nested workflows can be found in Appendix A. Finally, a small workflow for comparison of values represented in Figure 6 has been developed. This workflow performs a comparison of files and generates an SQL file for the registration of a new release of physical properties that the user may launch in the local DBMS engine if he deems it appropriate.

[4] http://www.taverna.org.uk

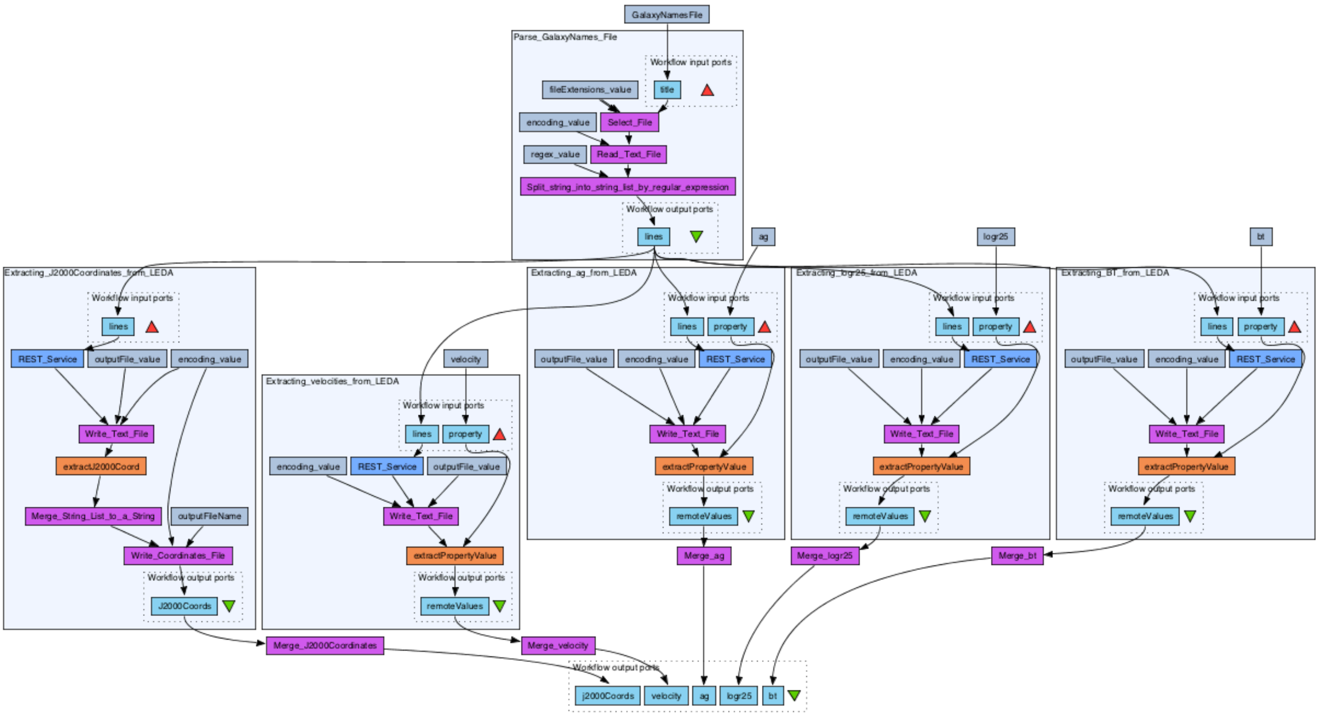


Figure 2: Workflow for gathering physical quantities from the HyperLEDA

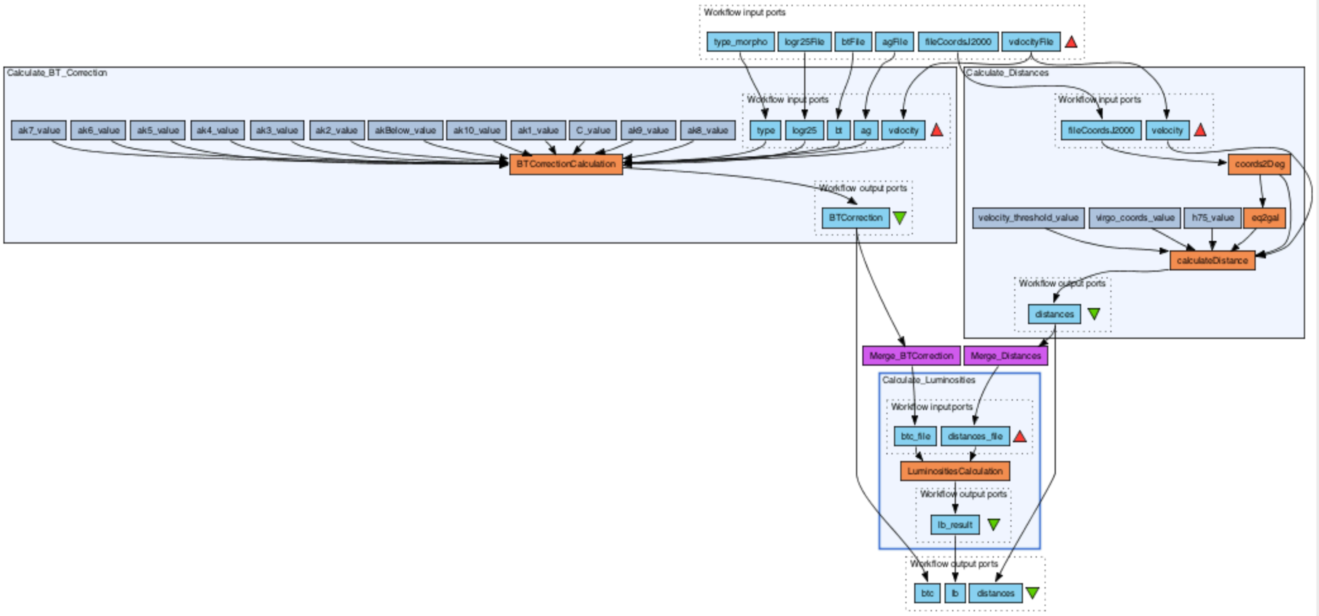


Figure 3: Workflow for propagation of physical quantities provided by the user

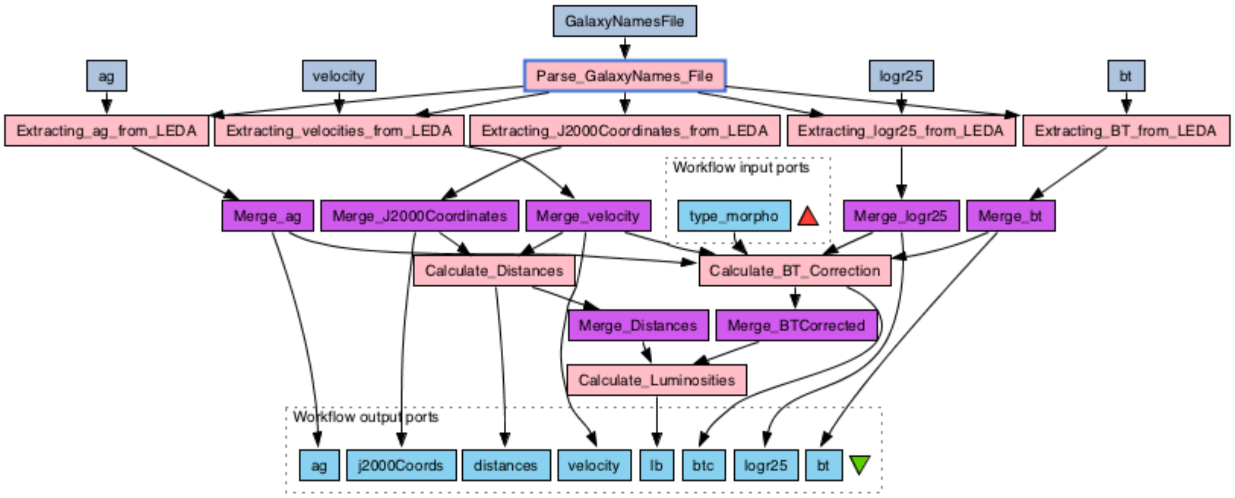


Figure 4: Workflow for calculation of luminosities with quantities coming from HyperLEDA

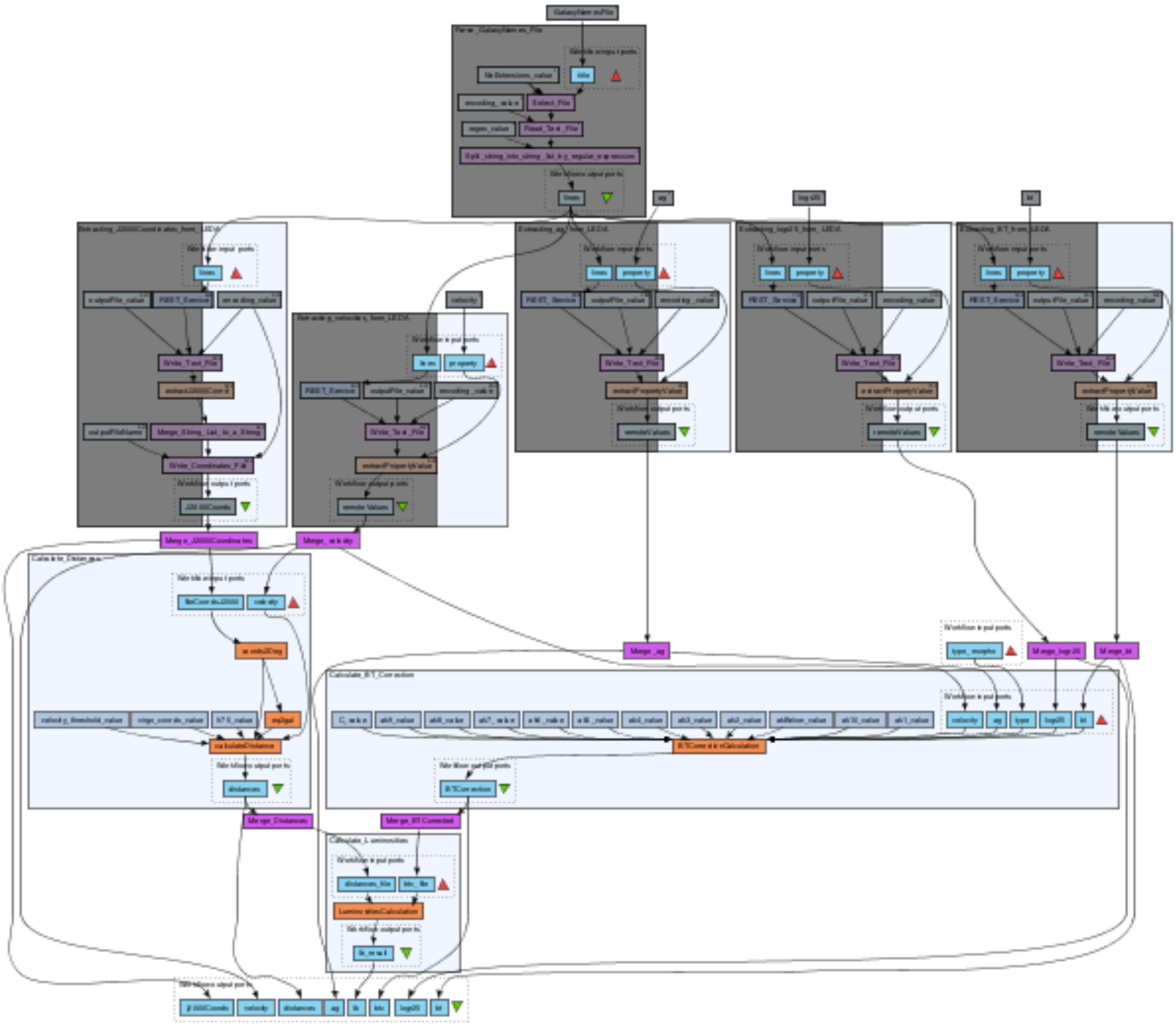


Figure 5: Running workflow for calculation of luminosities with quantities coming from HyperLEDA

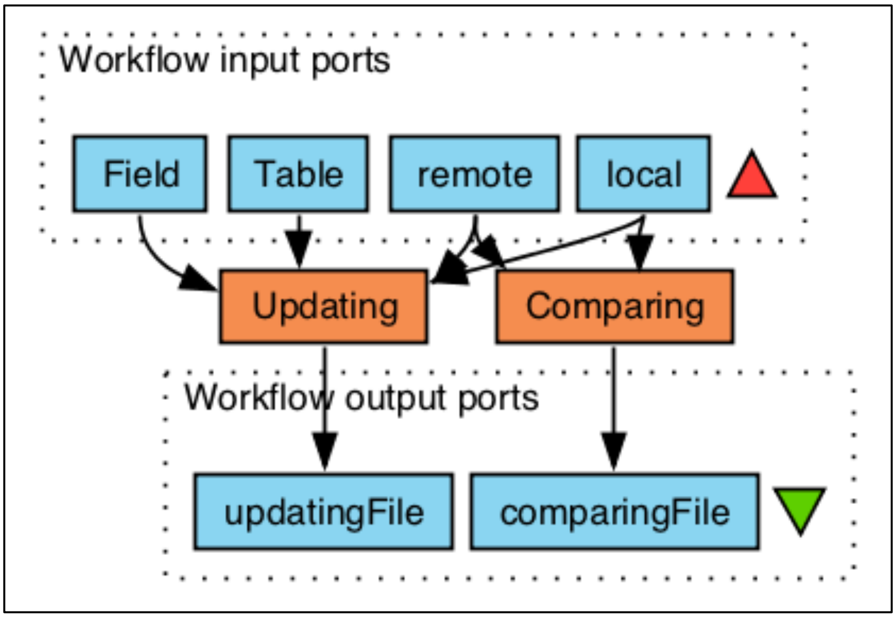


Figure 6: Workflow for comparison of physical quantities and update of user catalogue

2.5 Research Object Management

When dealing with this use case, the management of all the digital components involved in the process of curation of physical properties of AMIGA galaxies has been articulated through several tools developed in the Wf4Ever project. These tools have been provided to users of the incipient Wf4Ever platform as prototypes with the purpose of encouraging user feedback for the development of RO models and preservation services.

The ROBox [3] allows a seamless integration of the components of an RO in the Wf4Ever platform, where a shared folder in Dropbox[5] becomes a working RO. The content of the RO has been growing under a folder named *Quantities*, which is itself contained in the shared folder *AstroROs*. The detailed structure and content of the RO can be found in Appendix B.

The RO Manager Command Line Tool[6] has been used in order to annotate the components of the RO. A small extract of a working session with this tool can be found in Appendix C, where the commands used to configure the tool and synchronize with the Dropbox folder, create the RO *Quantities*, display its status, and annotate a workflow with several attributes as *type*, *keywords*, *description*, *format*, *note*, *title* and *creation date* can be found. All the information declared using this tool is registered in a RDF *manifest.rdf* file, which is included in the RO file structure. A sample of the content of this file can be found in Appendix C. Finally, the whole RO comprised in the *Quantities* folder has been zipped and uploaded to the MyExperiment[7] website as a *pack*.

[5] http://www.dropbox.com

[6] http://www.wf4ever-project.org/wiki/display/docs/RO+management+tool

[7] http://www.myexperiment.org/packs/231.html

3. Results

The results provided by the different workflows implemented in this use case are ASCII text files that contain values of physical properties for the 1051 AMIGA galaxies, differences between locally stored values and those provided by the HyperLEDA database or issued from calculations. Concerning the files of physical properties, they are formatted in a specific manner where each line contains a tabular separated data for one galaxy, which are the name of the galaxy, the actual data value for the physical property related to this file and the associated uncertainty. The files issued from the comparison workflows are SQL files and ASCII text files formatted in a similar way to the physical properties files. Those contain for each line, the name of the galaxy, the values for the actual data and uncertainties coming from the two comparing physical properties files, as well as the differences found in the data and uncertainties. Visual inspection of these results is needed in order to determine how the modification of certain parameters affect the final properties issued from calculations and the luminosities of galaxies in particular. The software Topcat[8] has been used to perform the inspection of the results. A set of plots representing the differences in the values vs. the identifiers of the galaxies (CIG[9] number) has been produced to ease the inspection tasks, those can be found in Appendix E.

3.1 Comparison of retrieved and final calculated physical properties

Differences between values of the local release dated on June 2010 and values provided by the HyperLEDA database in November 2011 are represented in the first seven figures in Appendix E. While there are no significant differences for the values of the dust extinction coefficient A_g (Figure 14) and neither for the values of axis ratio of the isophote 25 mag/arcsec^2 in the B-band *logr25* (Figure 16), more important differences can be found for values of the distances (Figure 20), velocities (Figure 17) and their uncertainties (Figure 18), and though in a minor extent also in values for the apparent total B magnitude B_T (Figure 15) and the corrected apparent B magnitude $m_{B\text{-}corr}$ (Figure 19). Since the order of magnitude in the differences for both B_T and $m_{B\text{-}corr}$ is very similar we can infer that differences in the parameters contributing to the correction of apparent total B magnitude are no significant. Finally, we can observe that differences in the intrinsic luminosities of galaxies (Figure 21) are quite significant for some galaxies.

3.2 Propagation of interdependent physical properties

The other figures in Appendix E represent how the final value of luminosities is affected when only one physical property is modified in the update of the HyperLEDA database, or how this property is propagated in the process of calculation of luminosities. It can be seen that the most important contributions come from differences in the velocities (Figure 25), in the apparent total B magnitude B_T (Figure 23) and in the morphological types (Figure 27). A unified representation on how the modifications on these three properties affect the final value of the luminosity can be found in Figure 28.

[8] http://www.star.bris.ac.uk/~mbt/topcat

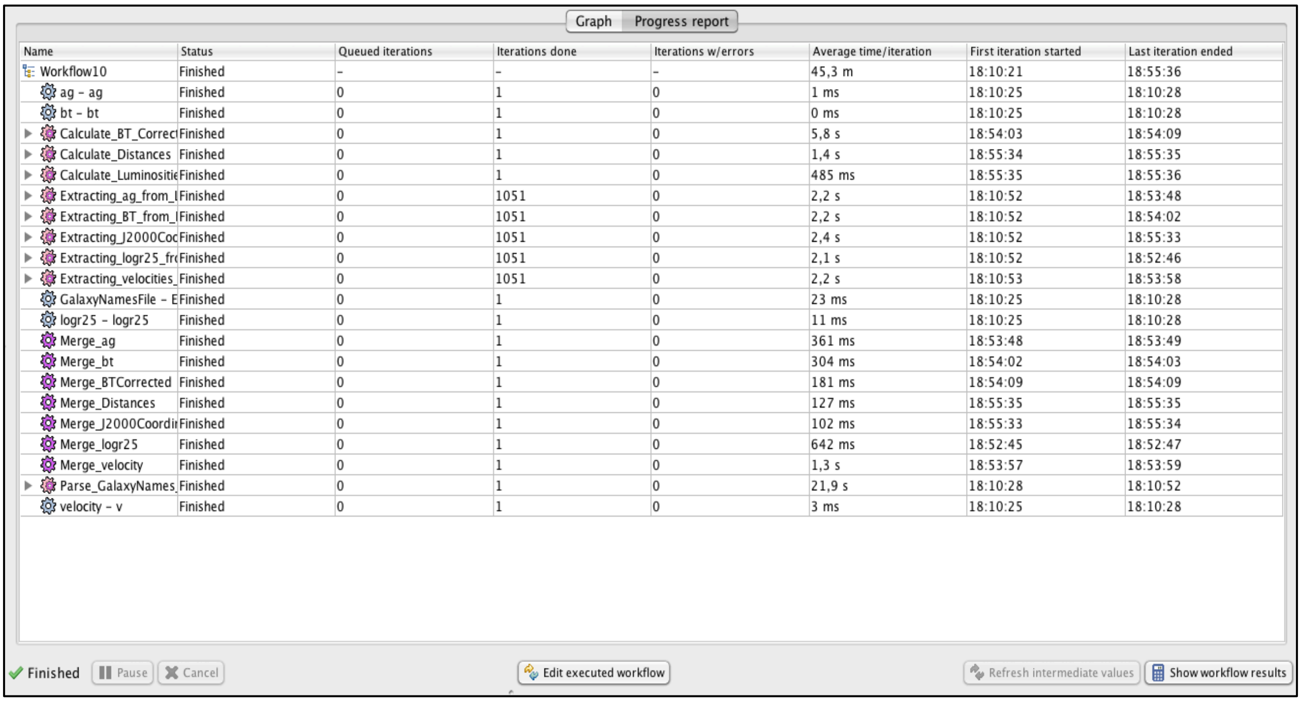
[9] Catalogue of Isolated Galaxies

4. Discussion

4.1 Impact

The curation tasks for a collection of values related to physical properties of galaxies and a preliminary study on how these are affected by the update of external databases has been confronted with a new workflow-based approach. The most significant impact can be seen in how the new methodology alleviates the complexity and allows an easy reproducibility of the process, provides processing modules that may be re-used in similar cross-boundary use cases dealing with triggered propagation of interdependent quantities and exposes the provenance of the entire process and highly valuable data.

We have studied the execution provenance of the workflow for the gathering of properties coming from HyperLEDA database and their propagation in the processes of calculation of luminosities. The performance in the execution of this workflow relies heavily on the time invested in the queries to HyperLEDA. The Figure 7 shows the *Progress report* of the execution of the workflow provided by Taverna. As it can be inferred from the values of the columns *First iteration started* and *Last iteration started*, most of the time in the execution is spent in the 1051 x 5 queries launched against the HyperLEDA web site. This performance could be improved since only one query per galaxy is needed for the extraction of five physical properties. Nevertheless, we have decided to split the query into five different queries and provide a more modular workflow, which is best suited for aims of re-usability and re-purposability. It may happen that for other physical properties, values may come from different databases.



Name	Status	Queued iterations	Iterations done	Iterations w/errors	Average time/iteration	First iteration started	Last iteration ended
Workflow10	Finished	-	-	-	45,3 m	18:10:21	18:55:36
ag - ag	Finished	0	1	0	1 ms	18:10:25	18:10:28
bt - bt	Finished	0	1	0	0 ms	18:10:25	18:10:28
Calculate_BT_Correct	Finished	0	1	0	5,8 s	18:54:03	18:54:09
Calculate_Distances	Finished	0	1	0	1,4 s	18:55:34	18:55:35
Calculate_Luminositie	Finished	0	1	0	485 ms	18:55:35	18:55:36
Extracting_ag_from_I	Finished	0	1051	0	2,2 s	18:10:52	18:53:48
Extracting_BT_from_I	Finished	0	1051	0	2,2 s	18:10:52	18:54:02
Extracting_J2000Coo	Finished	0	1051	0	2,4 s	18:10:52	18:55:33
Extracting_logr25_fr	Finished	0	1051	0	2,1 s	18:10:52	18:52:46
Extracting_velocities_	Finished	0	1051	0	2,2 s	18:10:53	18:53:58
GalaxyNamesFile - E	Finished	0	1	0	23 ms	18:10:25	18:10:28
logr25 - logr25	Finished	0	1	0	11 ms	18:10:25	18:10:28
Merge_ag	Finished	0	1	0	361 ms	18:53:48	18:53:49
Merge_bt	Finished	0	1	0	304 ms	18:54:02	18:54:03
Merge_BTCorrected	Finished	0	1	0	181 ms	18:54:09	18:54:09
Merge_Distances	Finished	0	1	0	127 ms	18:55:35	18:55:35
Merge_J2000Coordi	Finished	0	1	0	102 ms	18:55:33	18:55:34
Merge_logr25	Finished	0	1	0	642 ms	18:52:45	18:52:47
Merge_velocity	Finished	0	1	0	1,3 s	18:53:57	18:53:59
Parse_GalaxyNames_	Finished	0	1	0	21,9 s	18:10:28	18:10:52
velocity - v	Finished	0	1	0	3 ms	18:10:25	18:10:28

Figure 7: Progress report for the workflow of propagation of quantities obtained from HyperLEDA

4.2 Preservation and versioning

The case study covered in this document is particularly well suited to examine the initial requirements on preservation issues exposed in Wf4Ever delivered technical reports [1] [2] [3] [8] Since we are dealing with the curation of a user database that is affected by changes in an external database beyond the control of the user, he/she would need to be advertised of each of these external updates in order to be able to decide on the update of the local data, and the reason behind this update would also need to be preserved. It is interesting to note how these decisions are made depending on whether specific thresholds in the differences of values are exceeded, which brings the notion of conditional flow control logic in the process.

Updates in the external repository need to be versioned, which is not always the case currently. This would permit to keep a timeline of the evolution of these quantities: external experimental values, differences in these values between releases, local calculated values, differences in the values of local calculated values between releases, differences in the values of local values when calculated using a new value for a specific property, etc. These studies may allow the extraction of potentially relevant scientific information.

Information about the version of the RO and changes made in relation to previous versions would also need to be registered. When dealing with versioning, there are three major axes that contribute to make a new version of the RO without altering its identity. These are the complete set of *data* involved, the *processes* applied to the data and the *thresholds* used to make decisions. Any modification in one of these three axes would need to be documented, registered and preserved, in some cases providing the bibliographic reference on which the action has been based, and the authority of the user who made the change.

4.3 Annotations

Annotations on individual components of the RO seem to be an appropriate way to fill the potential lacks in the RO model as well as in the auto-generated information provided by future Wf4Ever tools or services. In this use case we consider that the most important annotations have been made using the Taverna workbench, where information for authoring, description and dataset examples has been provided for every workflow. In this sense the information related with dataset examples it is especially relevant, since it allows running a simple validation check of the workflow with the small exemplars input datasets provided as part of the annotations.

The RO Manager Command Line Tool has been used to make annotations on the components of the RO related to the attributes "*type*", "*keywords*", "*description*", "*format*", "*note*", "*title*" and "*creation date*", which may be extended in a future version of the tool. Other potentially useful attributes could be "*Additional info*" where the user could paste an URI to a resource providing additional information on the component. A "*Related to*" attribute could allow the linking of one or several others resources that the user consider related in some kind. Attributes for quality assessment could also be considered, as well as a "Similar" attribute where external web services or external data should be annotated in order to alleviate external resources decay.

Though it seems possible to annotate any single component in the RO, the functionality of managing annotations related to the whole RO could be really useful in order to register different general topics as the purpose of the research, still not well solved issues, assumptions made, hypothesis to be proven, etc. We have decided to register all this kind of information in three text files README.txt, RECIPE.txt and CONTENT.TXT at the root of the RO tree file structure. README.txt provides information concerning the purpose and identity of the RO, while RECIPE.txt tells the user how to use it, and CONTENT.txt presents the content of each folder of the RO. It could also be useful to improve the functionalities for annotations management, enabling edition and suppression of annotations, as well as the possibility to annotate folders.

4.4 Scenario for submission, dissemination and archival

A series of scenarios for the submission, archival and dissemination stages when working with the use case described in this document are proposed below. They can be seen as an attempt to determine the information needed for modelling a digital RO.

4.4.1 Submission

Alice has built a workflow for updating her catalogue of luminosities for a list of galaxies. She has used several Python scripts embedded in a Taverna workflow for calculating the final luminosities based on the values of apparent magnitudes, distances, galactic extinction, axis ratio and galactic coordinates provided by standard queries to the HyperLEDA database, but also on values that she provides for morphological types. The mathematical equations used in these scripts and information about the VO web services she has used are documented and discussed in already published literature (PDF papers). The workflow modifies the values of her catalogue of luminosities only if specific thresholds are exceeded at certain steps in the process of the propagation of updates. Some of the luminosities present a strange behaviour and Alice notes them in a text/log file. Alice thinks this workflow could be very useful for other users since it deals with a common problem: curation of physical properties of a list of celestial objects relying on external data that are subject to evolve due to improvements on sensor/instrument technologies. She would like to be credited for her work but she does not want some of her data to be shared.

When publishing her RO in the Wf4Ever platform, Alice provides information on what the RO intens to do with a small description of the workflows, the problems to solve and results or behaviour expected in other conditions. She aggregates to the RO a graphical and a digital representation of the workflows, the collection of incoming data, thresholds and scripts she has used, as well as the final calculated luminosities obtained and intermediate results provided by the services. A log file auto-generated in the execution of workflows exposes the provenance of the processes with all the queries, calculations and updates performed. This file is automatically aggregated with the digital representation of the workflow, as well as the date and time of its last enactment. She also adds some personal notes in a text/log file that provides information about specific weird values obtained or still not well solved issues.

She takes the time to manage restricted access on several of the components of the RO. The personal notes will be kept only for her while some of the incoming data, as the morphological types, may be accessed only from her group of colleagues, the rest is publicly accessible. She should provide some information about the services and data providers that have been used in the workflow (data release, date, instrument, waveband, filter, etc.) as well as some of the tags that better characterize and classifies her RO. She may publish later an updated version of the RO, declaring in which axis (incoming data, thresholds, processes) this new version of the workflow in the RO is different from the previous one.

4.4.2 Dissemination

Bob has a list of several tens of quasars that have been observed by members of his group during the last years. He is searching an RO that performs queries on VO repositories in order to gather additional complementary physical properties for his quasars. He has found the RO published by Alice browsing on the "Extragalactic" tag of the Wf4Ever platform. He reads the description of the workflow and finds some similarities to his problem; he also has a list of celestial objects and he would like to query an external on-line repository, though not the same as the one queried in Alice's RO. He had not thought about the possibility to calculate other more complex properties based on more basic ones provided by the VO, but now he finds it could be a good idea to calculate the intrinsic luminosity of his quasars as well as other properties in order to make some plots and look for some relations. He would have to find how to do it because the mathematical equations and quantities involved are not the same as those used in Alice's RO. He inspects some of the components of Alice's RO in order to understand it better and find those that he could re-use. He right-clicks on a file of velocities provided by a ConeSearch[10] VO web service and sends it to Topcat software via WebSAMP[11]. He inspect the file, sees the information on the columns, their semantic description, it seems that the VO service providing that file could be of great utility in his workflow. He looks for information related to that service that Alice has provided in the RO and the version of the data release. Bob decides to build a RO based on Alice's RO as a template.

4.4.3 Archival

The archived content of a RO should comprise workflows, input data and results, scripts used, published literature related to the experiment as well as other referenced digital entities, all the annotations the user has provided with annotation management tools (manifest.rdf file), and other small text files or logs accounting for possible specificities of the experiment. The structure and content of the RO considered in this document can be found in Appendix B. It may be interesting to note that the fact of considering several workflows affects the internal structure of the *data* folder, since output data of one workflow may be the input of another one. We have decided to keep as output all the data provided by workflows and as input only those provided by the user. Other data are *LB3D.txt*, a file built by the user with the help of external tools and *session.vot,* a session file of the Topcat software.

[10] http://www.ivoa.net/Documents/latest/ConeSearch.html

[11] http://www.star.bris.ac.uk/~mbt/websamp

5. Conclusions

The use of this RO as and advanced tool implies a real progress in the working methodology when the user is confronted to the Golden Exemplar. The time and efforts invested in the development and first implementation of the RO is similar to those spent with the previous handmade methodology used until now, with the significant added benefit of alleviating the complexity of the tasks and enabling their reproducibility. The graphical representation of the workflows allows a better understanding of the problem and the content of the RO is presented in a well-structured file organisation tree where components of different nature are annotated and thoroughly documented. Because of these benefits and because data curation is becoming a big need in astronomy research groups, we believe this RO represents a suitable use case for introducing the use of workflows and Wf4Ever RO working methodology to the astronomical community.

Since this Golden Exemplar is deeply related with versioning and preservation issues concerning local and external resources that are beyond the control of the user, the development of this RO has raised some considerations on when is necessary to store new versions of the RO or produce a completely different one, what differentiates versions from distinct RO identities, and what information has to be registered in the new version to keep track of the changes. Other issues pertain to access rights on RO components, what kind of data to be shared, how to share them and with whom, as well as considerations on what is needed to enable reuse and/or repurpose the RO. Other requirements inferred from the use of this RO are related to annotations, new attributes for the present annotations tool prototype have been proposed. The lack in the variety of attributes has been solved using text files (CONTENT.txt, README.txt, RECIPES.txt), where the content, purpose and functionalities of the RO have been described in a free format. These files are suitable to be a source of requirements for the annotations system.

6. References

[1] S. Bechhofer et al. *“Workflow Lifecycle Management Initial Requirements.”* Technical report, Deliverable 2.1, Wf4Ever project, 2011.

[2] R. González-Cabrero et al. *“Workflow Evolution, Sharing and Collaboration Initial Requirements.”* Technical report, Deliverable 3.1, Wf4Ever project, 2011.

[3] R. Palma et al. *“Wf4Ever Sandbox v1”.* Technical report, Deliverable 2.1, Wf4Ever project, 2011.

[4] G. Paturel et al. *“HYPERLEDA. I. Identification and designation of galaxies*.” Astronomy and Astrophysics, 2003, v.412, p.45-55.

[5] J. E. Ruiz et al. *“Virtual Observatory activities in the AMIGA group.”* Highlights of Spanish Astrophysics V, Astrophysics and Space Science Proceedings. Springer-Verlag Berlin Heidelberg, 2010, p. 533.

[6] I. Vauglin et al. *“Capabilities of the HYPERLEDA database.”* SF2A-2006: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics Eds.: D. Barret, F. Casoli, G. Lagache, A. Lecavelier, L. Pagani, 2006, p.365.

[7] L. Verdes-Montenegro et al. *“Astronomy Workflow Preservation Requirements.”* Technical report, Deliverable 5.1, Wf4Ever project, 2011.

[8] J. Zhao et al. *“Workflow Integrity and Authenticity Maintenance Initial Requirements.*“ Technical report, Deliverable 4.1, Wf4Ever project, 2011.

Appendix A – Graphical representation of nested workflows

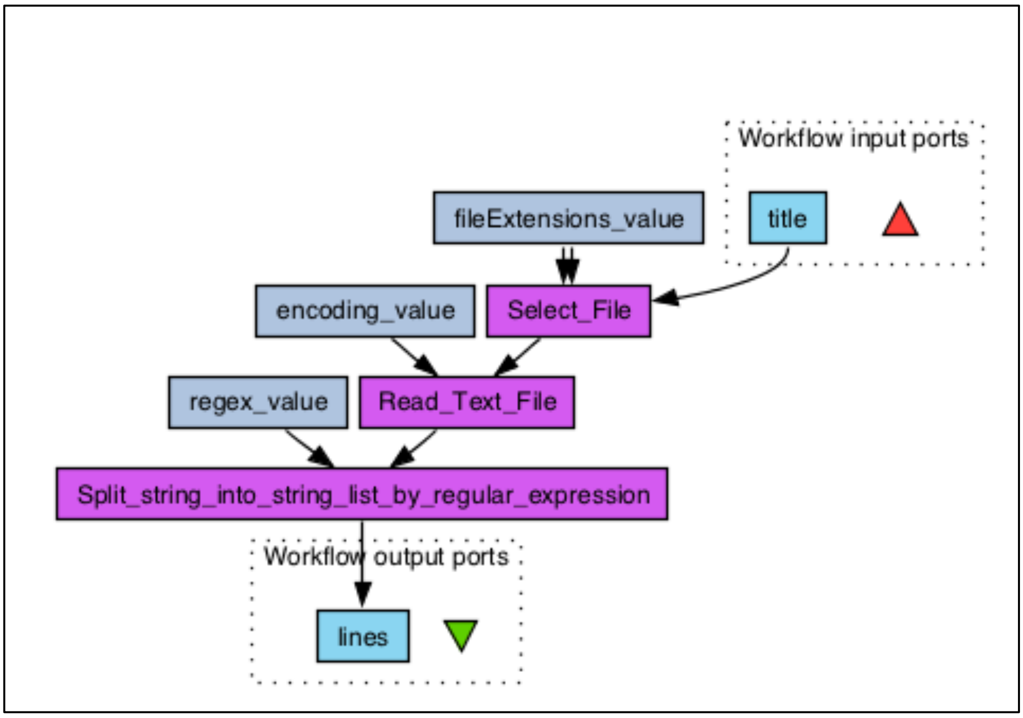


Figure 8: Workflow for parsing the GalaxyNames file

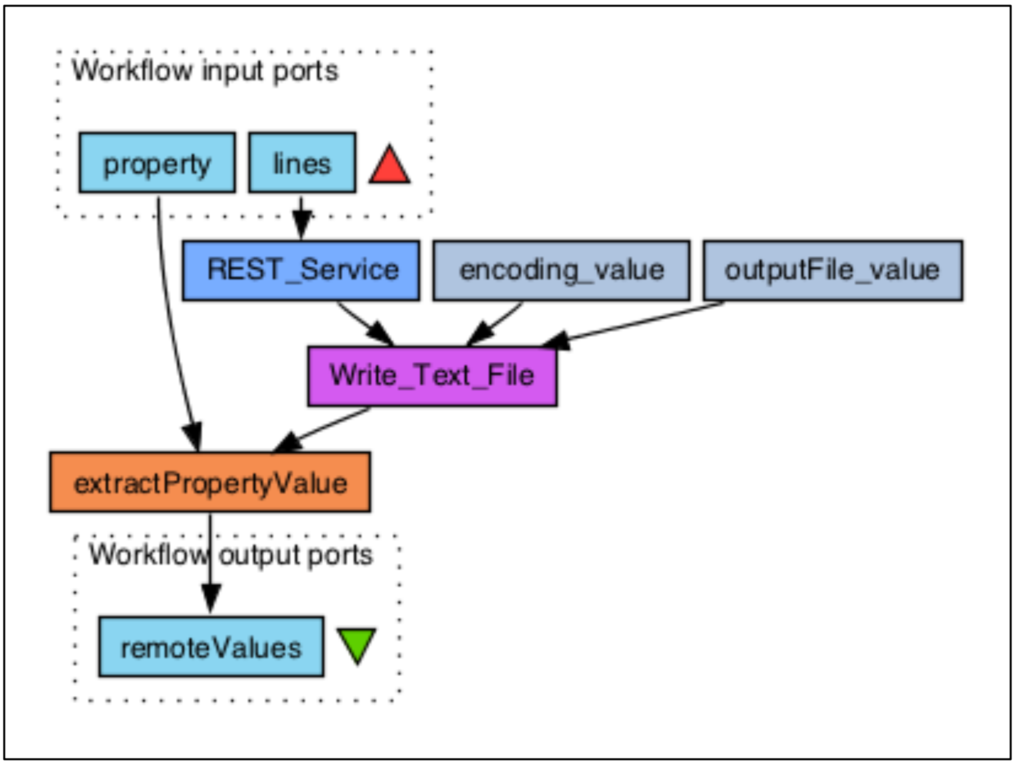


Figure 9: Workflow for extracting property values from HyperLEDA

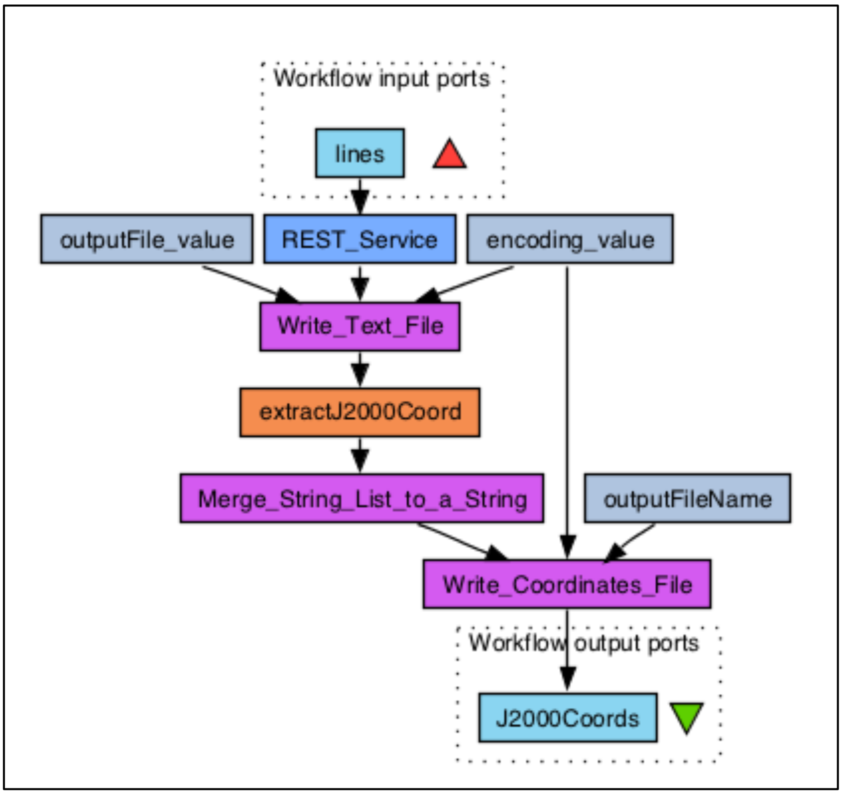


Figure 10: Workflow for extracting J2000 coordinates from HyperLEDA

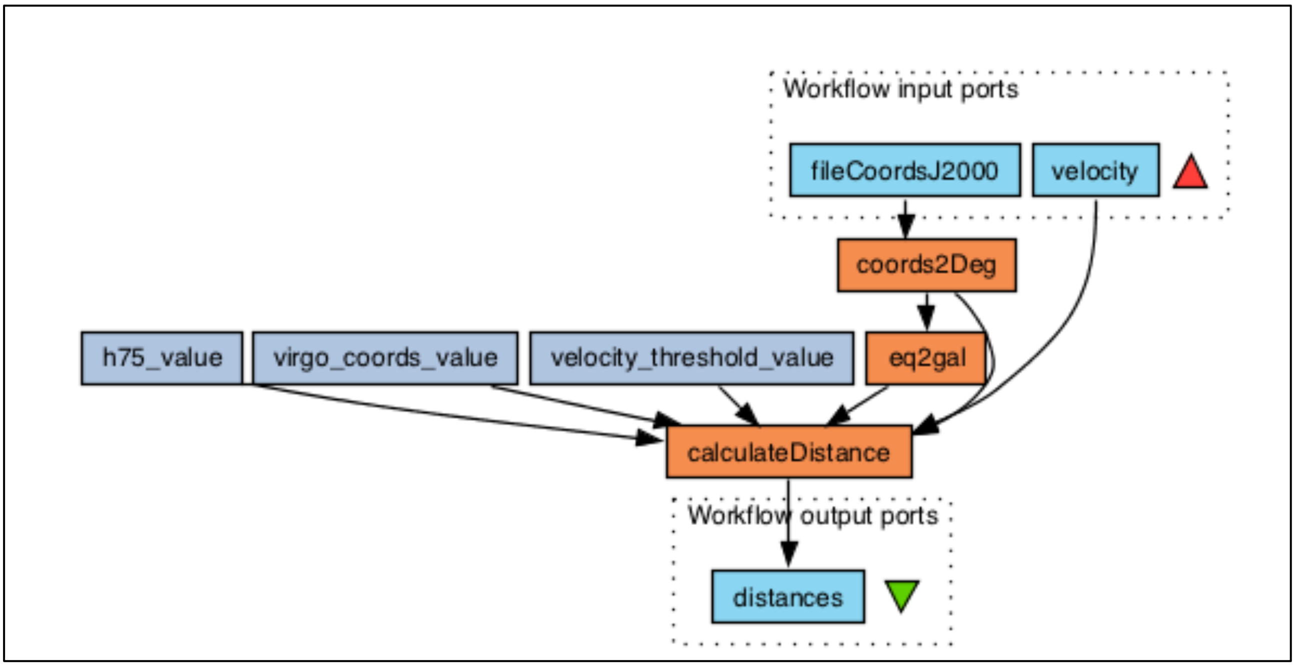


Figure 11: Workflow for calculating distances of galaxies

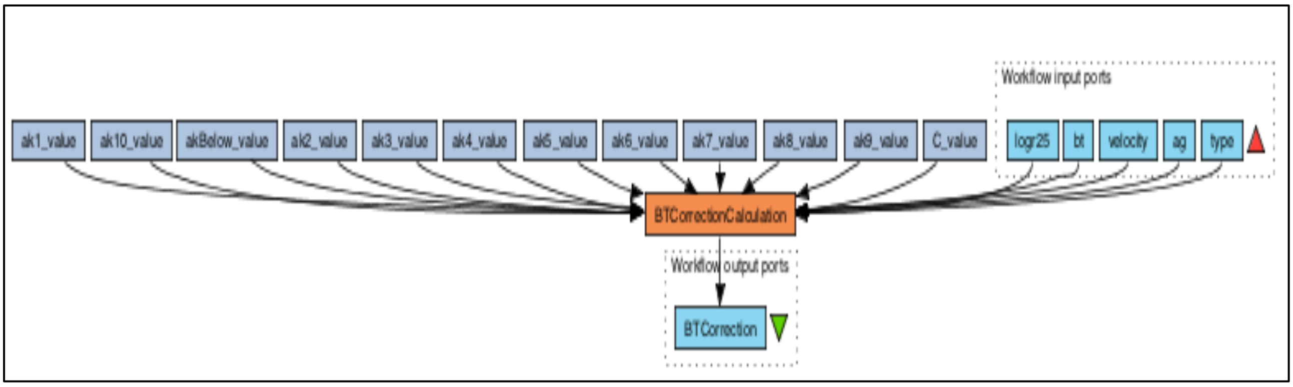


Figure 12: Workflow for calculating the correction for total B magnitude (BT)

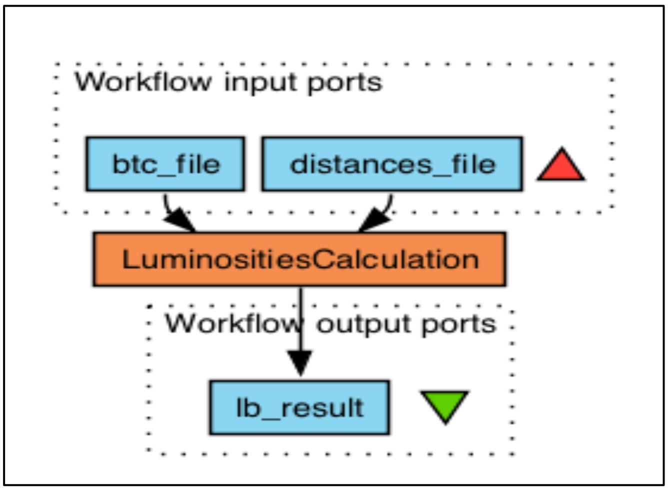


Figure 13: Workflow for calculating luminosities of galaxies

Appendix B – Research Object structure and content

```
|-Quantities
   |---.ro_manifest
   |---data
   |-----input
   |-------names
   |-------properties
   |-----others
   |-----output
   |-------diffs
   |-------properties
   |---references
   |-----bibliography
   |---scripts
   |-----comparing_updating
   |-----propagation
   |---sql
   |---workflows

./Quantities:
README.txt               RECIPE.txt                data                      references
scripts                  sql                       workflows

./Quantities/.ro_manifest:
manifest.rdf

./Quantities/data:
input                    others                    output

./Quantities/data/input:
names                    properties

./Quantities/data/input/names:
NamesLEDA.txt

./Quantities/data/input/properties:
agOld.txt                btcOld.txt                lbOld.txt                 morphoNew.txt
btOld.txt                distancesOld.txt          logr25Old.txt             morphoOld.txt
velocitiesOld.txt

./Quantities/data/others:
LB3d.txt                 session.vot

./Quantities/data/output:
diffs                     properties

./Quantities/data/output/diffs:
diff_ag.txt              diff_distances.txt        diff_lb_bt.txt            diff_lb_velocities.txt
diff_bt.txt              diff_lb.txt               diff_lb_logr25.txt        diff_logr25.txt
diff_btc.txt             diff_lb_ag.txt            diff_lb_morpho.txt        diff_morpho.txt
diff_velocities.txt

./Quantities/data/output/properties:
agNew.txt                distancesNew.txt          lb_ag.txt                 lb_morpho.txt
btNew.txt                j2000Coords.txt           lb_bt.txt                 lb_velocities.txt
btcNew.txt               lbNew.txt                 lb_logr25.txt             logr25New.txt
velocitiesNew.txt

./Quantities/references:
bibliography

./Quantities/references/bibliography:
D1.2.pdf                 D2.1.pdf                  D3.1.pdf                  D4.1.pdf
D5.1.pdf                 Paturel.pdf               Ruiz.pdf                  Vauglin.pdf

./Quantities/scripts:
comparing_updating       propagation

./Quantities/scripts/comparing_updating:
comparing.py             updating.py

./Quantities/scripts/propagation:
BTCorrectionCalc.py      coords2Deg.py             extractPropertyVal.py    luminositiesCalc.py
eq2gal.py                calculateDistance.py      extractJ2000Coord.py

./Quantities/sql:
ag.sql                    bt.sql                    btc.sql                  distances.sql
lb.sql                    logr25.sql                velocity.sql

./Quantities/workflows:
comparing_upd.t2flow     luminositiesVO.t2flow
gathering.t2flow         propagation.t2flow
```

Appendix C – RO Management Tool session commands

```
$ ROBASE = /Users/jer/Dropbox/AstroROs

$ ./ro help
Available commands are:

ro help
roconfig -b <robase> -r <roboxuri> -p <roboxpass> -u <username> -e <useremail>
ro create <RO-name> [ -d <dir> ] [ -i<RO-ident> ]
ro status [ -d <dir> ]
ro list [ -d <dir> ]
ro annotate <file><attribute-name> [ <attribute-value> ]
ro annotations [ <file> | -d <dir> ]

Supported annotation type names are:
type - Word or brief phrase describing type of Research Object component
keywords - List of key words or phrases associated with a Research Object component
description - Extended description of Research Object component
format - String indicating the data format of a Research Object component
note - String indicating notes on a Research Object component
title - Title of Research Object component
created - Date and time that Research Object component was created

See also:

ro --help

$ ./ro config
ROBOX service base directory:  /Users/jer/Dropbox/AstroROs/Quantities
URI for ROBOX service:         http://sandbox.wf4ever-
project.org/robox/dropbox_accounts/1/ro_containers/10
Password for ROBOX service:    d41d8cd98f00b204e9800998ecf8427e
Name of research object owner: Jose Enrique Ruiz
Email address of owner:        jer@iaa.es

$ ./ro create Quantities -d $ROBASE/Quantities -iQuantitiesId

$ ./ro status -d $ROBASE/Quantities
Research Object status
identifier:  QuantitiesId, title: Quantities
creator:     Jose Enrique Ruiz, created: 2011-11-17T10:36:08
path:        /Users/jer/Dropbox/AstroROs/Quantities
uri:         file:///Users/jer/Dropbox/AstroROs/Quantities/#
description: Quantities


$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow type "Workflow"
$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow keywords "LEDA, Extragalactic,
Properties, REST Services"
$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow description "Extraction of physical
quantities from LEDA catalog for a set of galaxies"
$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow format "Taverna 2.3"
$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow note "May be optimized if all calls
to LEDA are grouped"
$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow title "LEDA Extraction"
$ ./ro annotate $ROBASE/Quantities/workflows/gathering.t2flow created `date +"%Y-%m-%d"`


$ ./ro annotations /Users/jer/Dropbox/AstroROs/Quantities/workflows/gathering.t2flow
file:///Users/jer/Dropbox/AstroROs/Quantities/workflows/gathering.t2flow
title: LEDA Extraction
format: Taverna 2.3
keywords: LEDA, Galaxies, Properties, REST Services
created: 2011-11-17
note: May be optimized if all calls to LEDA are grouped
description: Extraction of physical quantities from LEDA catalog for a set of galaxies
type: Workflow
```

Appendix D – Manifest file for Research Object

```
<rdf:RDF
xml:base=".."
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:ns1="http://ro.example.org/ro/terms/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
<rdf:Descriptionrdf:about="workflows/gathering.t2flow">
<dcterms:description>Extraction of physical quantities from LEDA catalog for 1051 isolated AMIGA
galaxies</dcterms:description>
<dcterms:title>LEDA Extraction</dcterms:title>
<ns1:note>May be optimized if all calls to LEDA are grouped</ns1:note>
<dcterms:created>2011-11-17</dcterms:created>
<dcterms:type>Workflow</dcterms:type>
<dcterms:format>Taverna 2.3</dcterms:format>
<dcterms:subject>AMIGA, LEDA, Extragalactic, Properties, REST Services</dcterms:subject>
</rdf:Description>
<rdf:Descriptionrdf:about="workflows/README.txt">
<ns1:note>Content of RO in .ro_manifest/manifest.rdf</ns1:note>
</rdf:Description>
<rdf:Descriptionrdf:about="workflows/propagation.t2flow">
<dcterms:description>Calculation of luminosities of galaxies with data provided in ASCII
files</dcterms:description>
<dcterms:title>Luminosity calculation</dcterms:title>
<ns1:note>Data files to provide may be ASCII files in a very specific format</ns1:note>
<dcterms:created>2011-11-17</dcterms:created>
<dcterms:type>Workflow</dcterms:type>
<dcterms:format>Taverna 2.3</dcterms:format>
<dcterms:subject>Galaxies, Luminosity, Calculation</dcterms:subject>
</rdf:Description>
<rdf:Descriptionrdf:about="workflows/comparing_updating.t2flow">
<dcterms:description>Comparison of data files and generation of SQL files for insertion on new
tables</dcterms:description>
<dcterms:title>Datafiles comparison and SQL insertion</dcterms:title>
<ns1:note>Data files to compare may be ASCII files in a very specific format</ns1:note>
<dcterms:created>2011-11-17</dcterms:created>
<dcterms:type>Workflow</dcterms:type>
<dcterms:format>Taverna 2.3</dcterms:format>
<dcterms:subject>Datafiles, Databases, Comparison, Insertion</dcterms:subject>
</rdf:Description>
<rdf:Descriptionrdf:about="README.txt">
<dcterms:description>Description of the digital experiment and Research Object</dcterms:description>
<dcterms:title>RO Description</dcterms:title>
<dcterms:created>2011-11-17</dcterms:created>
<dcterms:type>Document</dcterms:type>
<dcterms:format>ASCII</dcterms:format>
<dcterms:subject>Description, Abstract, Hypothesis</dcterms:subject>
</rdf:Description>
<rdf:Descriptionrdf:about="#">
<dcterms:description>Quantities</dcterms:description>
<dcterms:title>Quantities</dcterms:title>
<dcterms:created>2011-11-17T10:36:08</dcterms:created>
<rdf:typerdf:resource="http://vocab.ox.ac.uk/dataset/schema#Grouping"/>
<dcterms:creator>Jose Enrique Ruiz</dcterms:creator>
<dcterms:identifier>QuantitiesId</dcterms:identifier>
</rdf:Description>
<rdf:Descriptionrdf:about="RECIPE.txt">
<dcterms:description>Description of the methodology, how to use this Research Object</dcterms:description>
<dcterms:title>RO Methods</dcterms:title>
<dcterms:created>2011-11-17</dcterms:created>
<dcterms:type>Document</dcterms:type>
<dcterms:format>ASCII</dcterms:format>
<dcterms:subject>Description, Methods</dcterms:subject>
</rdf:Description>
<rdf:Descriptionrdf:about="workflows/luminositiesVO.t2flow">
<dcterms:description>Calculation of luminosities of galaxies with data extracted from the LEDA
catalog</dcterms:description>
<dcterms:title>Luminosity calculation with LEDA properties extraction</dcterms:title>
<ns1:note>May be optimized if all calls to LEDA are grouped</ns1:note>
<ns1:note>Content of RO in .ro_manifest/manifest.rdf</ns1:note>
<dcterms:created>2011-11-17</dcterms:created>
<dcterms:type>Workflow</dcterms:type>
<dcterms:format>Taverna 2.3</dcterms:format>
<dcterms:subject>LEDA, Galaxies, Properties, Luminosity, Calculation</dcterms:subject>
</rdf:Description>
</rdf:RDF>
```

Appendix E – Plots of workflows results

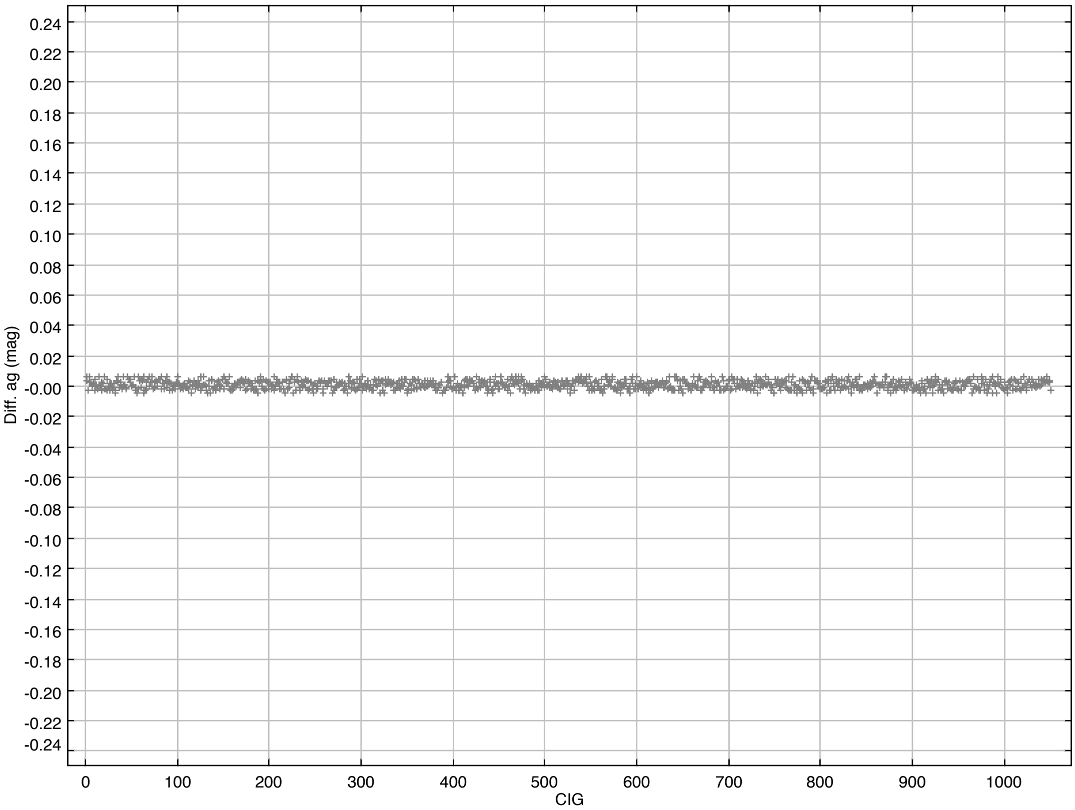


Figure 14: Differences in the galactic extinction (Ag) in the comparison process

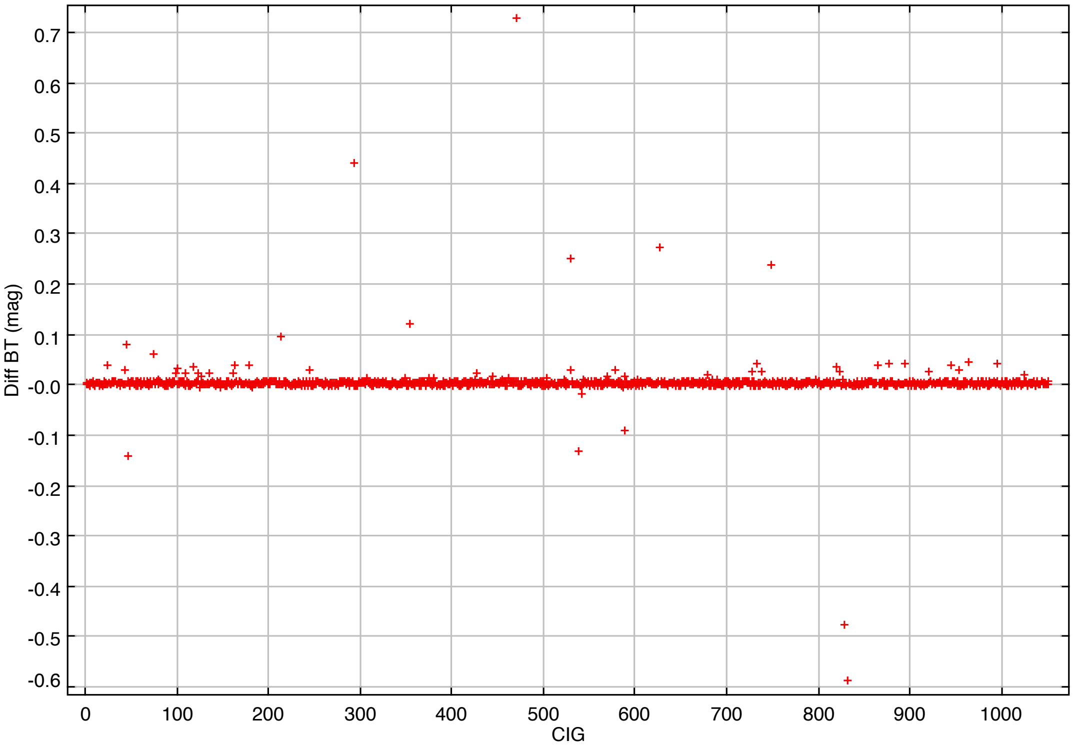


Figure 15: Differences in the total B magnitude (BT) in the comparison process

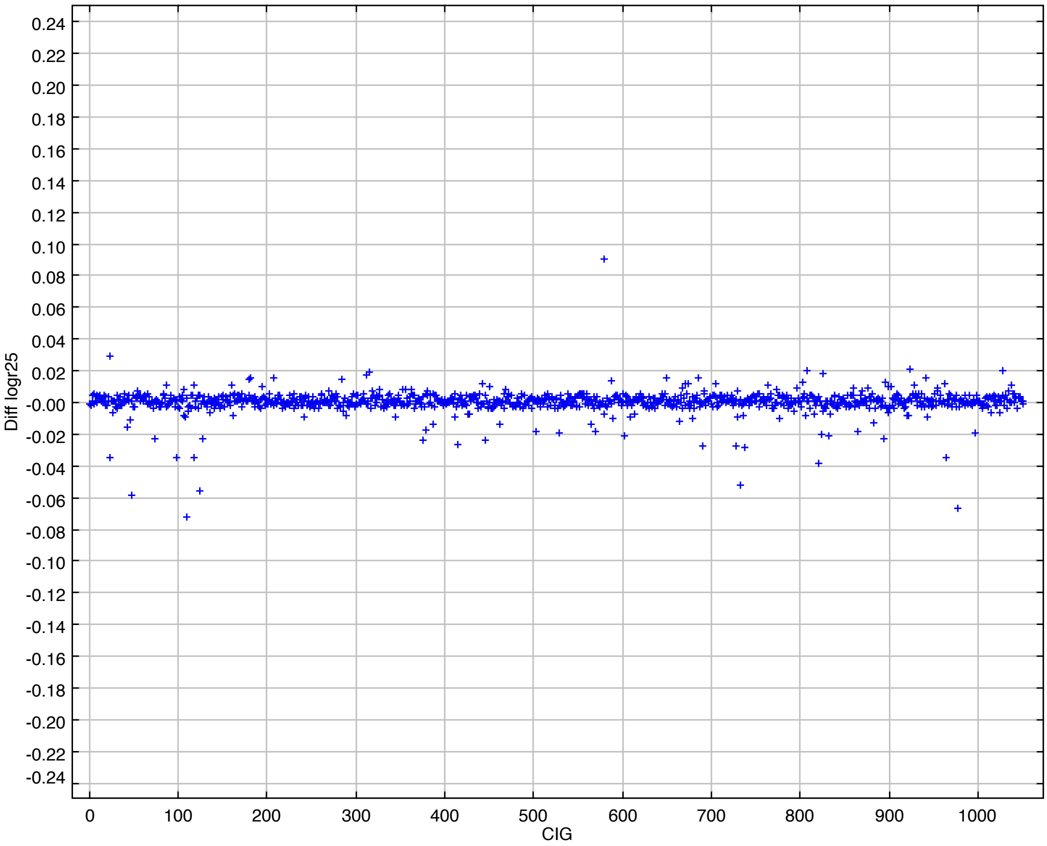


Figure 16: Differences in the log of axis ratio (logr25) in the comparison process

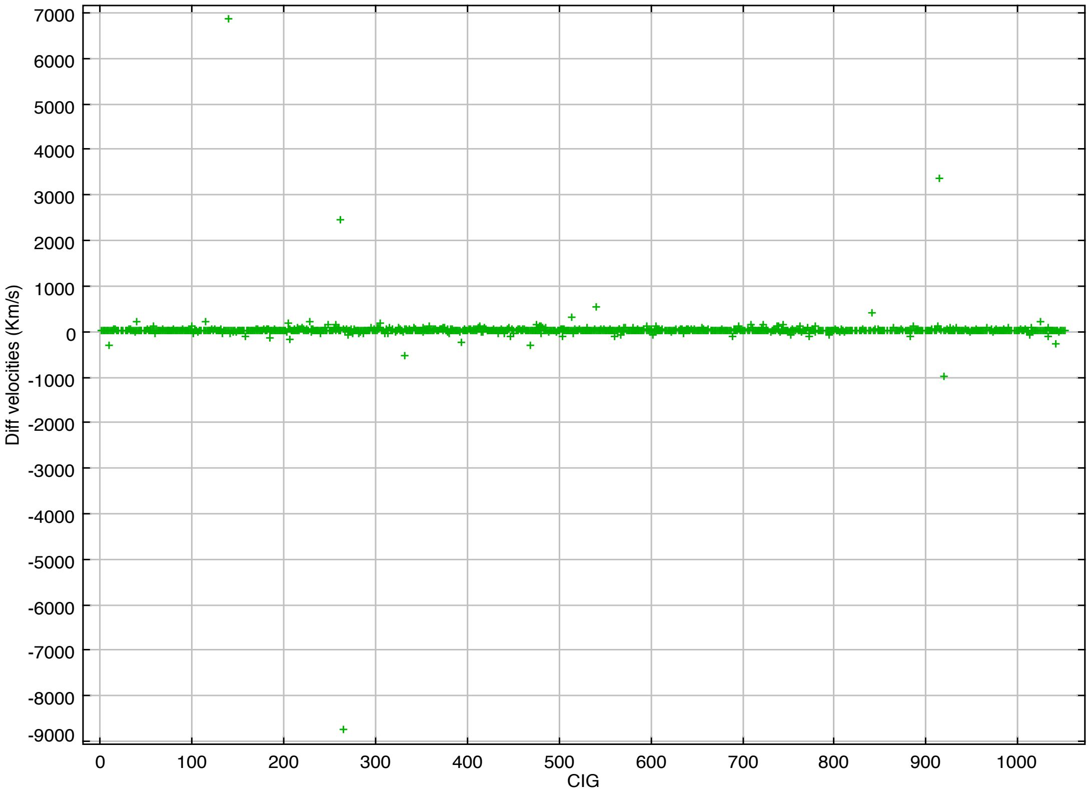


Figure 17: Differences in the radial velocities in the comparison process

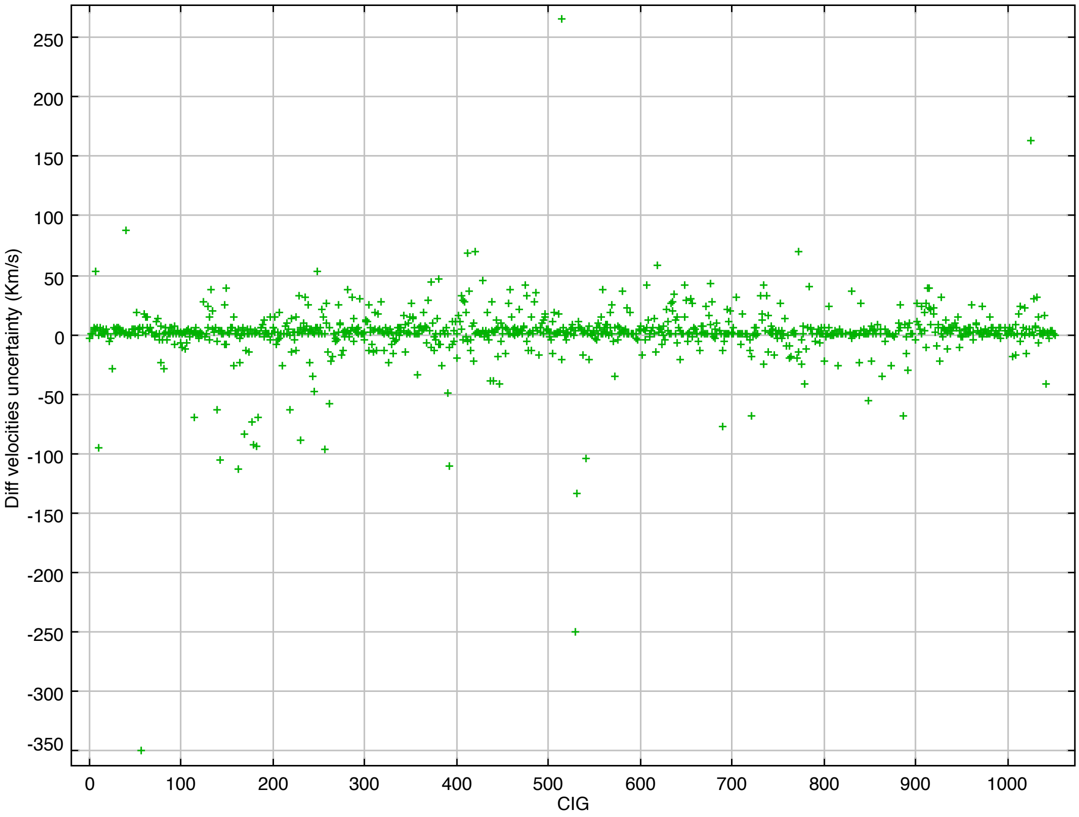


Figure 18: Differences in the uncertainty of the radial velocities in the comparison process

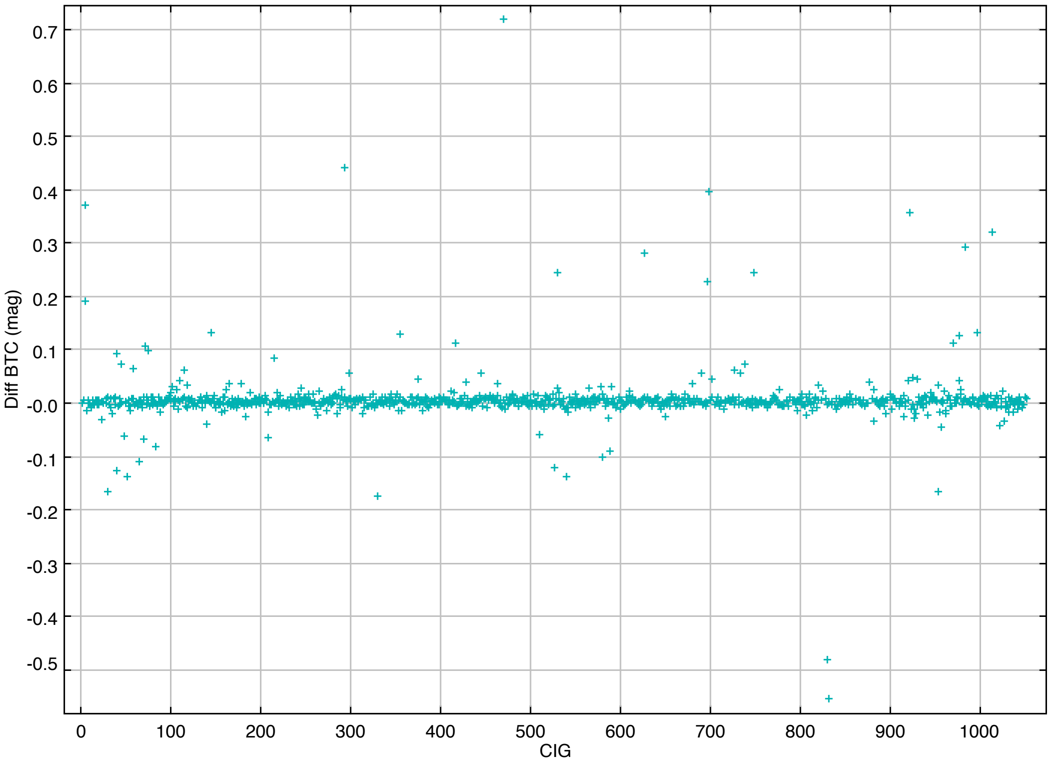


Figure 19: Differences in the total corrected B magnitude (mB-corr) in the comparison process

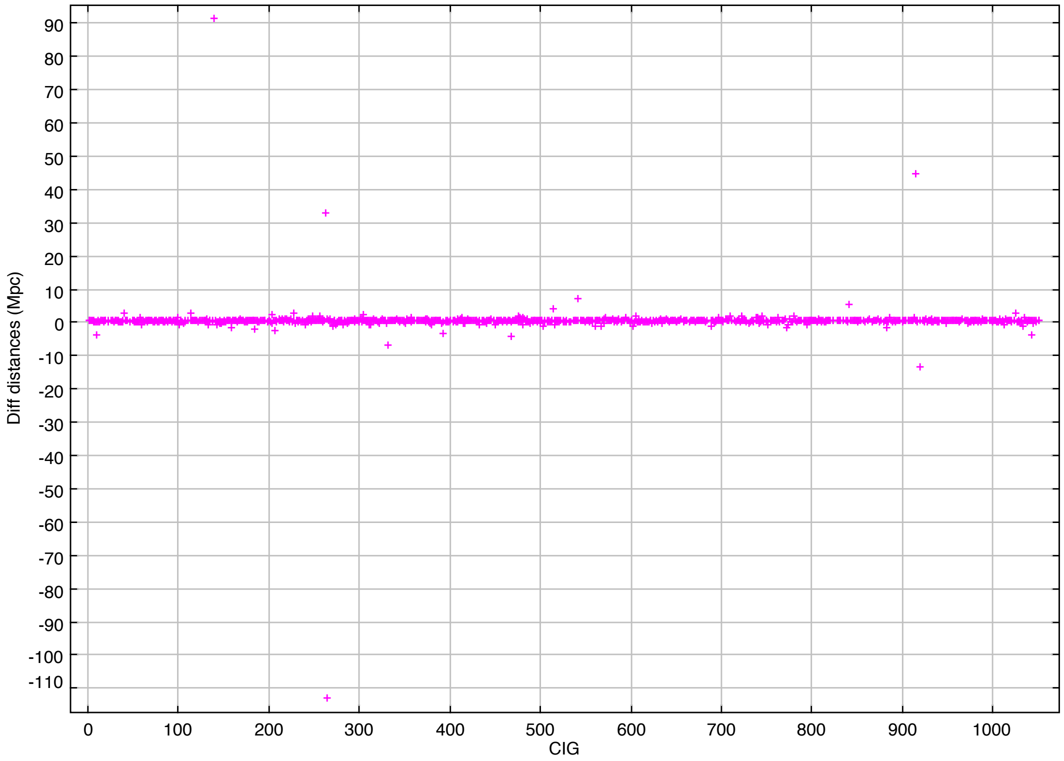


Figure 20: Differences in the distances in the comparison process

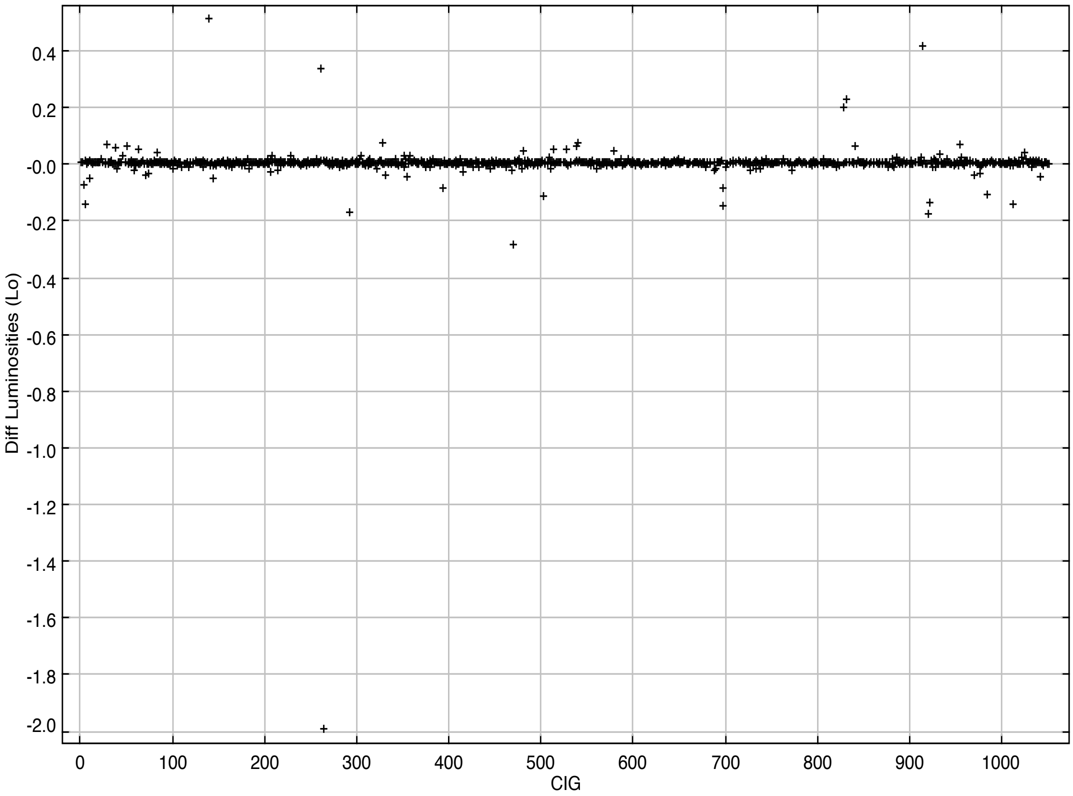


Figure 21: Differences in the luminosities in the comparison process

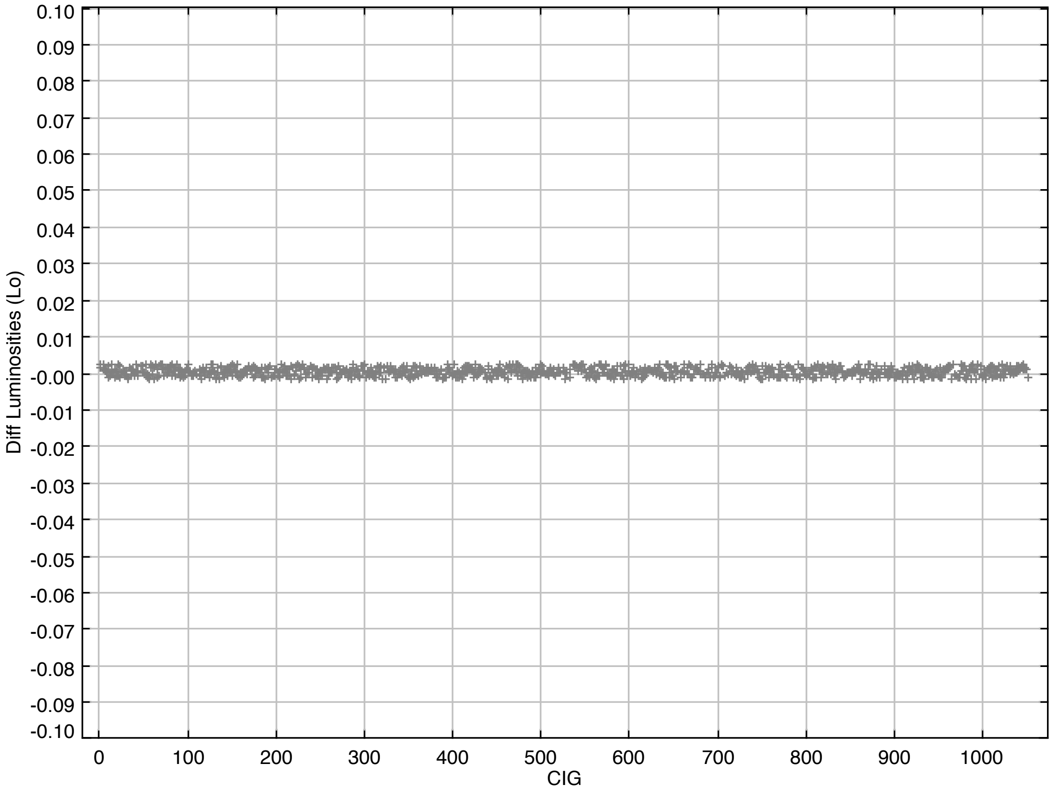


Figure 22: Differences in the luminosities when only affected by galactic extinction (Ag)

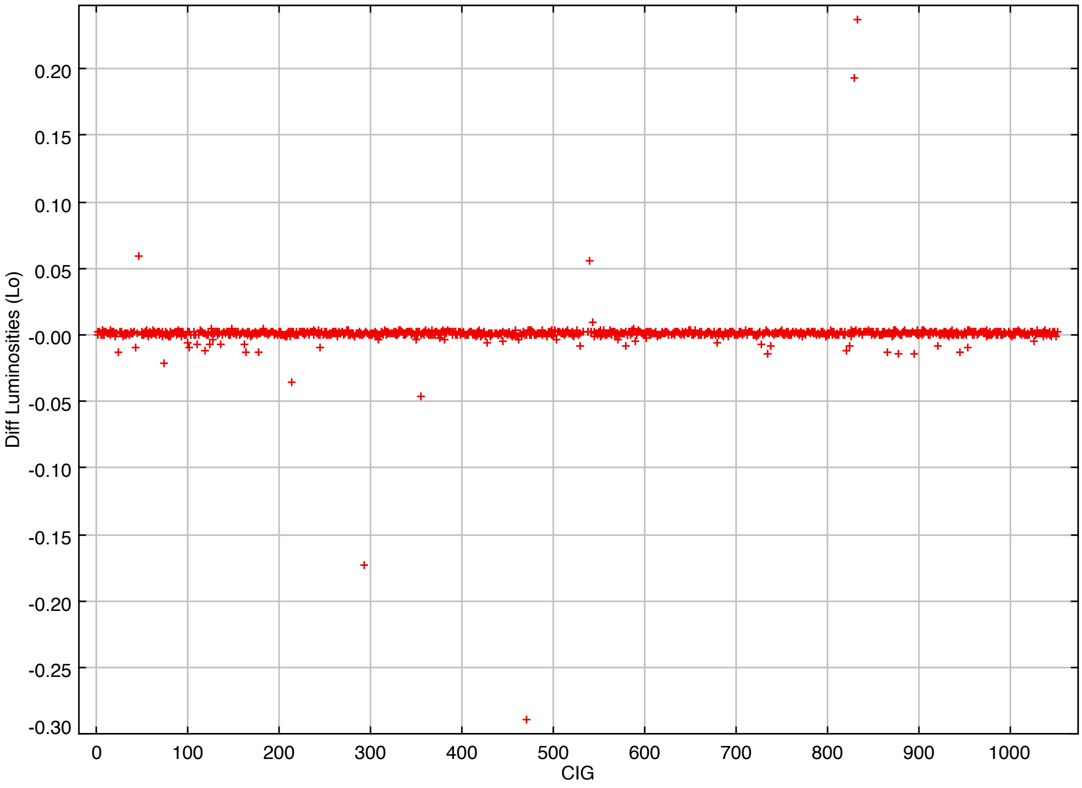


Figure 23: Differences in the luminosities when only affected by total B magnitude (BT)

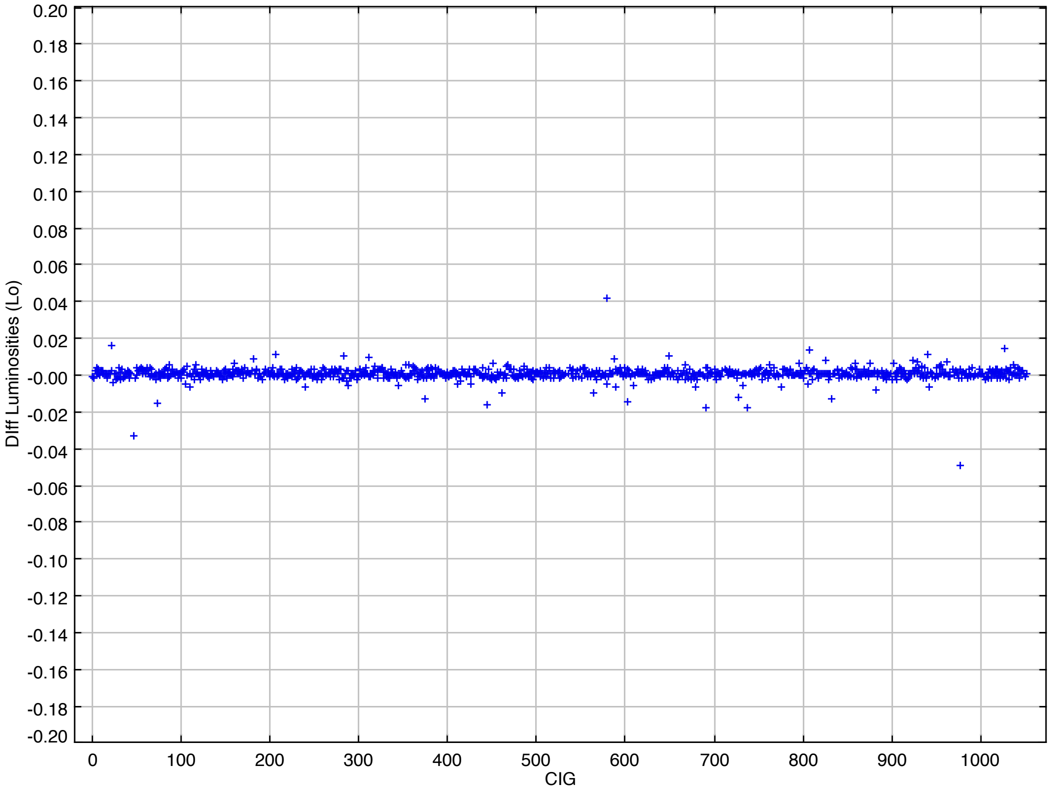


Figure 24: Differences in the luminosities when only affected by the log of axis ratio (logr25)

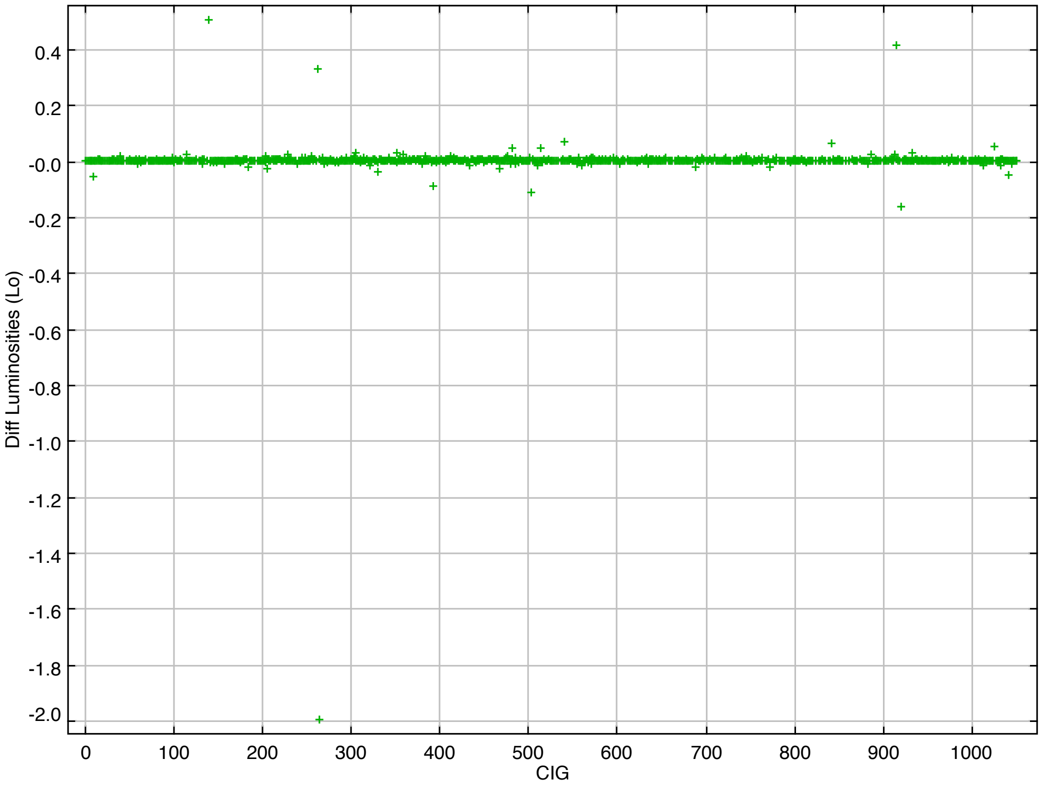


Figure 25: Differences in the luminosities when only affected by the velocities

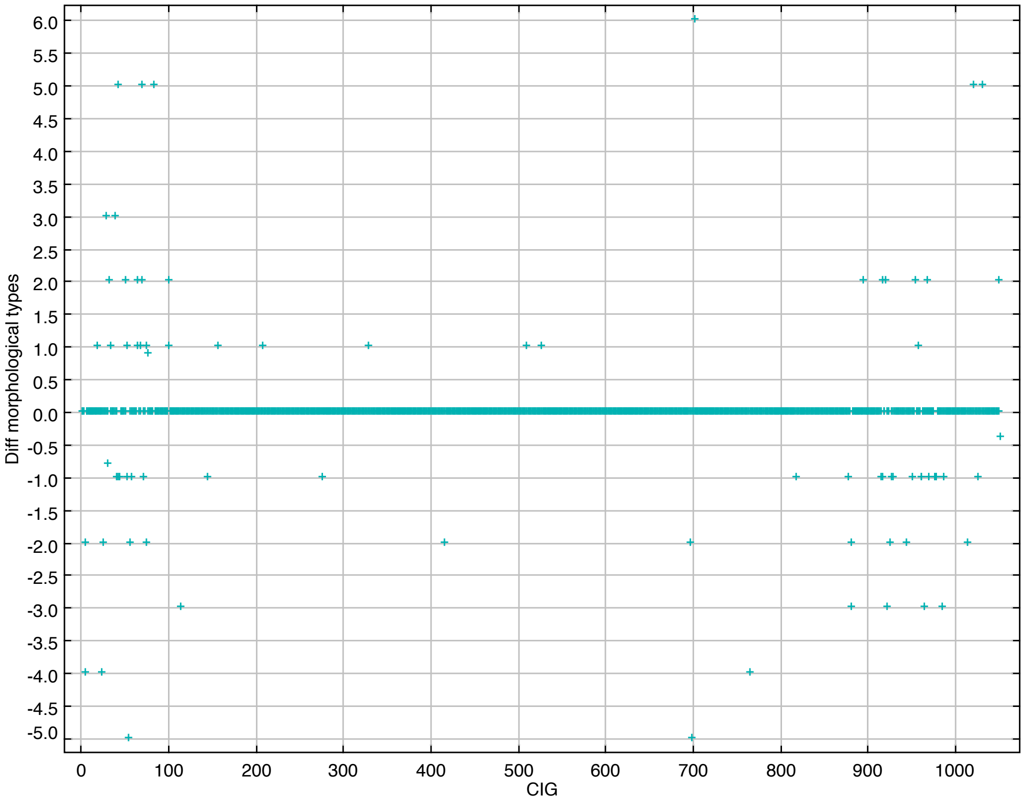


Figure 26: Differences in the morphological types used

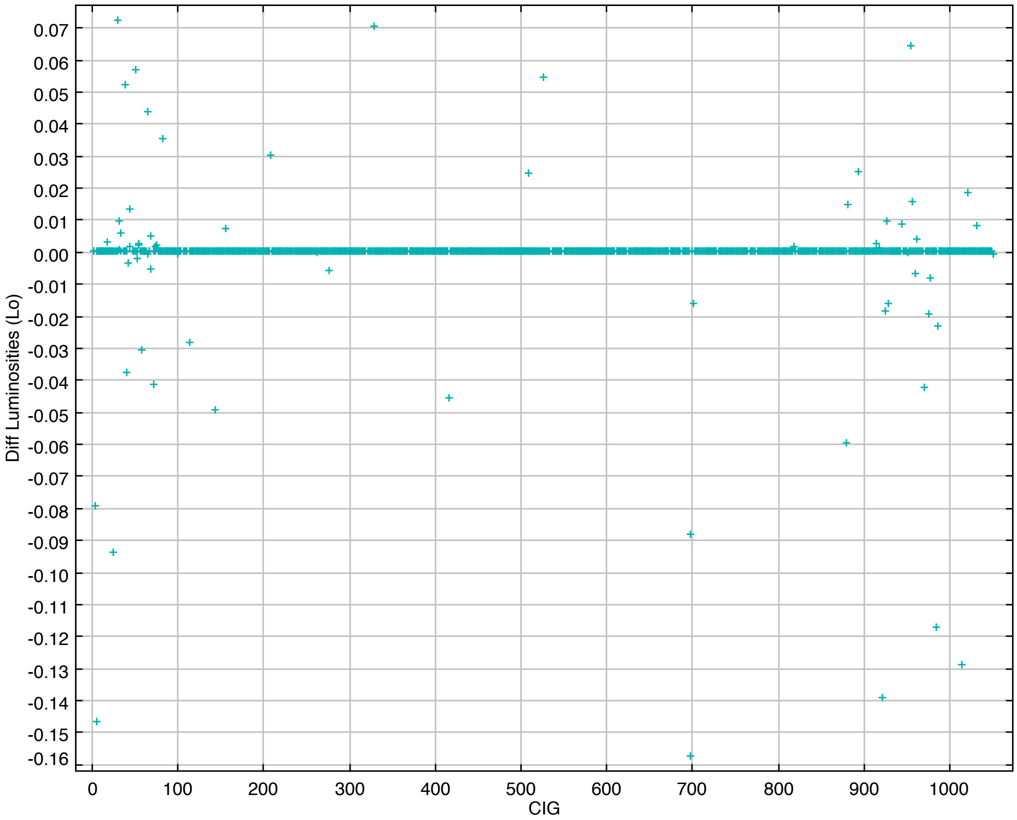


Figure 27: Differences in the luminosities when only affected by morphological types

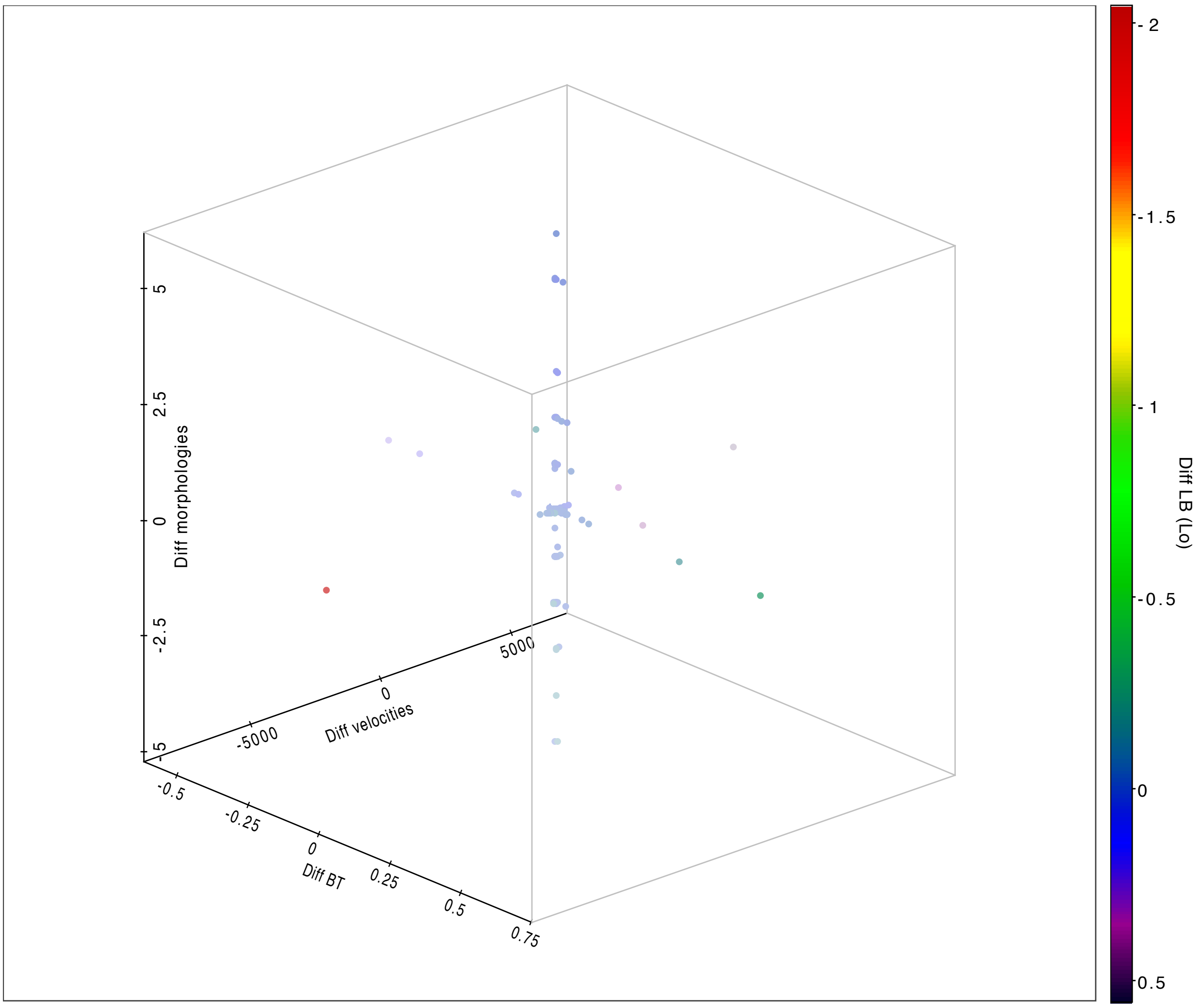


Figure 28: Differences in the luminosities w.r.t. differences in other quantities