Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science

STREP FP7-ICT-2007-6 270192

Objective ICT-2009.4.1 b) – "Advanced preservation scenarios"

D5.3v2: Calculation of Luminosity Profiles for a Sample of Galaxies extracted from Catalogues using Isolation Criteria

---

Deliverable Co-ordinator: José Enrique Ruiz (IAA)

Deliverable Co-ordinating Institution: IAA

Other Authors: Dr. Lourdes Verdes-Montenegro (IAA); Julián Garrido (IAA); Susana Sánchez (IAA); Dr. Mirian Fernández (IAA); Dr. José Sabater (Univ. Edinburgh); Dr. Juan de Dios Santander (IAA)

Document Identifier:	Wf4ever/2010/D5.3v2/v0.2	Date due:	30/09/2012
Class Deliverable:	Wf4ever 270192	Submission date:	**30/09/2012**
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	11-09-2012	José Enrique Ruiz	Initial Draft
0.2	13-09-2012	José Enrique Ruiz; Lourdes Verdes-Montenegro; Julián Garrido; Susana Sánchez; Mirian Fernández; José Sabater	Additions on Sections “Materials and Methods”
0.3	14-09-2012	José Enrique Ruiz	Additions on Section “Discussion”
0.4	15-09-2012	José Enrique Ruiz	General revision and minor additions
0.5	17-09-2012	José Enrique Ruiz	Final Draft before QA
0.6	28-09-2012	José Enrique Ruiz	Corrections derived from QA
1.0	30-09-2012	Lourdes Verdes-Montenegro; José Enrique Ruiz	Final state

Executive Summary

This document describes the development of the second Golden Exemplar proposed in Work Package 5: Workflow Astronomy Preservation, and provides discussion about user-driven issues risen in this process and related with general objectives of the Wf4Ever project. The document pertains to the assembly of two Research Objects (RO) packing the digital experiment undertaken for the extraction of a sample of galaxies based on environmental criteria and the calculation of luminosity profiles in several bands for each of the galaxies. The main purpose of deliverable D5.3v2 is to produce the workflows and ROs for the second Golden Exemplar in order to provide feedback based on a user experience for the models and technologies developed in the Wf4Ever project. The workflows and ROs developed may be accessed publicly in the MyExperiment[1] portal, where the two ROs produced have been uploaded as MyExperiment Packs, as well as in the RODL Wf4Ever Sandbox[2] [4] [5] .

The scientific experiment represented is related with the extraction of a sample of galaxies from a database provided by public archives as well as its photometric study in several observed bands. Two different approaches have been considered in the development of these workflows and ROs: the migration of already existing Python scripts into Taverna workflows for the characterization of the target sample constitutes the strategy followed in the development of the first RO, while in the second one we decided to start from scratch the conception and development of the workflows related to the photometric study. The fact of considering a large sample of galaxies and the complexity of the whole protocol to follow makes this use case suitable for a methodology benefiting from a workflow automated and reproducible approach. Furthermore, the work implies collaboration between different experts, and when the results published, will allow other researchers to perform the study using the same methodology with another sample, so that a comparison with galaxies in other environments (pairs, groups) is valid.

We provide a detailed description of the Golden Exemplar including information about the implementation process of the ROs and the scientific results obtained after the enactment of the workflows. We also propose a generic tree-folder structure to register and expose the many different relationships among the components of a RO in the file system, as well as their role and nature, since we still do not have the tools to visualize them as semantic web annotations provided in the RO building process. Finally, we provide a discussion about how the use of these ROs and a set of best practices is already having an impact on the present working methodology undertaken in the team, taking also into account preservation and versioning issues as well as quality and completeness checking of the experiment.

[1] http://www.myexperiment.org

[2] http://sandbox.wf4ever-project.org/portal

Table of contents

List of Figures

1. Introduction

The formation and evolution of galaxies is related to the environment and the interactions with other galaxies. The properties of a galaxy depend on its history of interactions with companion galaxies. Galaxies located in dense groups or clusters are usually more massive than isolated galaxies and often show an elliptical shape (instead of spiral). On the other hand, galaxies that are interacting usually show signs of recent star formation and a higher prevalence of active supermassive black holes.

Hence, in order to understand the evolution of galaxies, detailed analysis of astronomical images plays a key role. During the last years, telescopes working at different wavelengths have been imaging the whole sky. One example of this is the Sloan Digital Sky Survey (SDSS, see Sect 2.1), whose generated images provide a wealth of information, publicly available, and hence allowing them to be used by any interested group in the study of their own samples of objects or generation of new samples. This is a powerful tool for open science and an outpost in the Big Data era. The two main kind of processing performed on the images generated by sky surveys are explained below.

Extraction of sources: point and extended objects can be found in images of the sky, most frequently the first corresponding to stars, and the second to galaxies. Although their visual identification is in general straightforward, production of samples of objects according to different criteria becomes not only more complex, but inefficient to be done in an interactive non-automated visual process. For this reason the first RO of the 2nd GE of the Astronomy use case is focused on tasks to generate samples of objects with specific characteristics.

Structural analysis of the light for each source: once a target source has been identified, the astronomer would perform different types of analysis to determine its characteristics, as measuring the total light or size, its radial distribution, or 2D shape. In the case of galaxies, the last could correspond to determining the number of spiral arms, presence of a bar, ratio between the bulge and disk component, etc. and modelling of them. The second RO of the 2nd GE works on extracted sources identified as galaxies in the previous steps, obtaining the radial profile of the light, as well as fitting of bulge and disk components.

The result of this study will be applied by the AMIGA group of astronomers of the Wf4Ever project in order to determine the properties of isolated galaxies, those who spend most of their life in absence of strong interactions with neighbours, and test the hypothesis that for this sample the bulge component has a smaller size than for other interacting samples. This would be expected if bulges grow, as models predict, due to accretion of material from interactions. Such study requires the analysis of a large sample so that it has a statistical value. However the implied processes are complex enough to prevent for an efficient work when they have to be executed manually and the protocol is not thoroughly described.

A description of the used methodology, resources and Wf4Ever tools is provided in Sect. 2 and the scientific results are summarized in Sect. 3. A discussion about how this developments impact on present research and working methodology on the AMIGA group, as well as issues risen in the RO-ification process related with tools, semantic annotations, quality and preservation can be found in Sect. 4. Sect. 5 is dedicated to the conclusions.

2. Materials and methods

2.1 The SLOAN Digital Survey

The Sloan Digital Sky Survey[3] (SDSS) is an ambitious project that consists on observing more than a quarter of the sky using the 2.5-meter telescope at Apache Point Observatory, New Mexico. Since the data collection began in 2000, it has obtained multi-colour images of around 500 million objects, and spectra for more than 1 million objects. SDSS data have been released to the scientific community and the general public in annual increments, with the final public data release from SDSS occurring in August 2012. All the images, catalogues of measurements, and spectra are accessible online through web forms and web services querying efficient databases, and a lot of tools have been designed to browse through sky images, look up data for individual objects, or search for objects anywhere in the sky based on any criteria. SDSS data actually supports an enormous range of scientific investigations by astronomers around the world, and have become one of the main tools in astronomy research.

2.2 Extraction and characterization of a sample of galaxies

A careful quantification of the environment and interactions of a galaxy is fundamental to draw conclusions about its formation and evolution. A correlation between these environmental parameters and the properties of galaxies can explain which of the latter ones are intrinsic to the galaxy (nature) and which are induced by the interaction with companions (nurture). However this is not an easy task, since the quantification of the environment requires knowing the position of the companion galaxies to elucidate their association with the target galaxy. It is not possible to accurately measure the real position of a galaxy in the 3D space, but its projected position on the sky, and this makes very difficult to check whether two nearby galaxies are really associated or not. The use of spectroscopy combined with the Hubble law allow to estimate the distance to a galaxy (with some error).

An approach to tackle the nature vs. nurture topic is to use robust estimators of the environment and tidal forces (interaction with companions) in a large sample of galaxies. The SDSS provides an ideal catalog to estimate these derived properties. Galaxies having a redshift between 0.03 and 0.1 covered by the SDSS spectroscopic survey compose the sample to be studied by this first RO. A catalogue of potential nearby companions is obtained for each target galaxy. The properties of the potential companions were obtained from the SDSS photometric survey public archive. The developed workflows evaluate in a first step the real or fake association of a potential companion with the target galaxy, based on its optical properties (magnitude in r-band, distance to the target galaxy, spectroscopic redshift and photometric redshift). After that, two environmental estimators are derived for each target galaxy, taking into account only the companions that have been inferred to be really physically associated. The first parameter accounts for the density of companions of similar mass in a region around the target galaxy and the second parameter estimates the gravitational forces exerted by these companions. The results allow the study of the correlation of these parameters with other properties of the galaxies.

[3] http://www.sdss.org

2.3 Calculation of luminosity profiles

The study of galaxy properties is deeply related with the electromagnetic energy that is emitted and absorbed by their different components. Thus, looking at different regions of the spectrum, it is possible to study each of them. Using optical images and spectra the stellar population age, the chemical composition of the stars, their distribution within the galaxy, and even their internal motions can be studied.

To study the distribution of stars within the galaxy, we use the luminosity profiles, which give information on the distribution of the stars as a function of the distance to the centre of the galaxy. In order to extract these profiles, an optical image of the galaxy (calibrated in physical units) is needed. From this image, we measure the surface brightness of the galaxy at different distances from the galaxy centre, so allowing the identification of different structural components, such as bulge, disk, or presence of a bar. Quantification of the shape of the shape of these profiles provides information about how the galaxy formed and evolved.

Different tools can be used to obtain the information described above. In this second RO we have used two complementary software packages: while GALFIT makes a 2-dimensional modelling of the galaxy and calculates the luminosity profiles of the modelled components, ELLIPSE extracts the luminosity profile of the observed galaxy. The comparison between modelled and observed luminosity profiles will allow us to extract the relevant information for our study.

2.4 Photometry modelling software

2.4.1 SExtractor

SExtractor[4] (Source-Extractor) is a software package that builds a catalogue of objects extracted from an astronomical image. In order to produce this catalogue of structural and photometric values for each object identified in the image, SExtractor needs as an input configuration file with some information about the image. This configuration file also contains a large number of control parameters that will be used for the identification of objects, including a minimum area, a sigma level above the sky background, and a deblending parameter for separating overlapping objects. In our specific case, we are only interested in the largest object of the image, hence the default value is used for all these fine-tuning.

2.4.2 GALFIT

GALFIT[5] is a data analysis algorithm that fits 2-D analytic functions to the brightness distribution of galaxies registered in digital images. The used functions are such as an exponential, a Sérsic profile or a de Vaucouleurs profile, which are widele found in the astronomical literature,. One of the main features of GALFIT is that it allows for the simultaneous fitting of an arbitrary number of structural components, as well as combinations of the above functional forms.

[4] http://www.astromatic.net/software/sextractor

[5] http://users.obs.carnegiescience.edu/peng/work/galfit/galfit.html

One of the scientific applications of GALFIT is to model large, spatially resolved galaxies, in order to probe their detailed structures. With the aim of studying the nearby isolated galaxies of the AMIGA[6] sample, we have designed a workflow to make a bulge-disk-bar decomposition applying GALFIT to images provided by the SDSS DR7[7]. Since GALFIT needs some initial parameters, the workflow is designed to take them from the results produced by SExtractor, via a previous workflow. The results of GALFIT include a tabular file with the fitted parameters of each structural component as well as a multi-layer FITS file composed of 2-D images including *observed galaxy*, *modelled*, and *residuals* images, which can also reveal evidence of other fine structures.

2.4.3 ELLIPSE IRAF

The ELLIPSE task works under the IRAF[8] environment and is used to fit the isophotes (equal brightness curves) of optical images of galaxies with elliptical curves. The task reads a 2D image and produces as main output one tabular file containing the parameters of each fitted isophote. The initial values come from a first guess provided by the user of approximate values for the centre, ellipticity and position angle of the galaxy. We have designed a workflow that takes these initial values from the results produced by a previous workflow using SExtractor. Then both the modelled images provided by GALFIT and the observed ones are fitted. The isophotes are fitted in increasing values of the radii following a pre-defined rule, the fit for a previously fitted ellipse being used as an input for the next one. Using the table given by ELLIPSE, we end by plotting the values representing the luminosity profile of the galaxy.

2.5 Execution environment

The first requirement to run the workflows composing the two considered ROs is Taverna Workbench[9] 2.4 or higher. AstroTaverna[10] (Taverna plugin) is also needed in order to get functionalities related with Virtual Observatory web services queries and management of standard VOTable data formats.

In general, the execution environment is a Linux distribution including Python[11] 2.x and a bash shell. Access to a PostgreSQL[12] database storing the physical parameters provided by SDSS is also needed; a dump file of the database may be downloaded from the AMIGA web server in order to be deployed and accessible from a local execution environment. The following list shows other specific dependencies for the main workflows:

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

[7] http://www.sdss.org/dr7

[8] http://iraf.noao.edu

[9] http://www.taverna.org.uk/download/workbench/2-4/

[10] http://wf4ever.github.com/astrotaverna/

[11] http://www.python.org/

[12] http://www.postgresql.org/

- Initialize
  - Python with *psycopg*[13] package
- Environment
  - Python with *psycopg* and *numpy*[14] package
- Preprocessing
  - Python
  - read_PSF[15]
  - AstroTaverna
- SExtractor
  - SExtractor 2.8.6 or higher
  - AstroTaverna
- Galfit
  - Galfit
  - AstroTaverna
- Ellipse
  - IRAF[16] including *tables*, *stsdas* and *ellipse* modules
  - Python with *pyraf*[17] package
  - AstroTaverna
- Plots
  - Python including *matplotlib*[18] package
  - AstroTaverna

2.6 Strategy and workflow building

We decided to split the Golden Exemplar in two different ROs: Sample Selection and Luminosity Profiles. The first one generates a sample of galaxies characterizing their environment, and the second one fits models to the images of the galaxies and extracts their luminosity profiles. The accomplished modularity allow for other potential complementary studies of the target sample, as well as its application to other samples. Different approaches have been also undertaken for the development of the ROs. The environment characterization has been designed as a migration of pre-existing Python scripts to the workflow methodology, focusing on the automation of tasks for a large sample of galaxies, making intensive use of a database engine and avoiding as much as possible local software dependencies. The photometric study has

[13] http://initd.org/psycopg/docs/

[14] http://numpy.scipy.org/

[15] http://www.sdss.org/dr7/products/images/read_psf.html

[16] http://www.astro.uson.mx/~favilac/downloads/ubuntu-iraf/iso/IRAF_Ubuntu.iso

[17] http://www.stsci.edu/institute/software_hardware/pyraf

[18] http://matplotlib.org/

been designed as a migration of human actions and procedures present on the experimental protocol, focusing on the transparency of the methodology and tasks for a small sample of galaxies, and making use of local standard commonly used software very well known by the astronomer.

We have developed our workflows so that the outcome of a specific one is the income of the following, many intermediate products not needed in subsequent ones. In the workflow design we have been forced to take decisions on how to manage error exceptions in order have flow process not interrupted by eventual software crashes found for some galaxies of the sample. We have made very intensive use of VOTable standards admonished by IVOA as interoperable format for astronomical data management, as well as SDSS VO Service. These VO-related actions are much more easily handled with the AstroTaverna plugin developed in the Wf4Ever project. The fact of working with local software for some of the workflows requires some input parameters to be PATHS and filenames in the local file system. This kind of information is provided by the user as configuration files, well differentiated from scientific related information.

Sometimes, the need of scripting has forced us to code small Python scripts that are stored “by value” in the t2flow Taverna workflow file. Although Taverna offers the possibility to provide them “by reference” as external URL accessible files, we decided to keep the code in the workflow in order to provide the user with a unified working environment (i.e. Taverna workbench) for the development and modification of workflows. Finally, we would like to stress on our effort to manage error exceptions for generic designed workflows. The need of different software configuration files for each galaxy of the sample lead us to consider the creation of these files on-the-fly from generic templates and a tabular file provided by the user. A generic template-filler needs the association of each one of these data to their correspondent blanks in the templates. The workflows guarantee their internal consistency by checking that the terms used in the templates and the tables are the same, and moreover that they also match a default vocabulary file.

2.6.1 Sample Selection. Initialize

This workflow saves a tabular *.*pckl* Python pickle dataset in the local file system, containing values calculated on physical parameters associated to potential companions of a sample of target galaxies. These original physical parameters are extracted from a postgreSQL database, containing information of all galaxies covered by the SDSS spectroscopic survey. The workflow first access the external database located in the AMIGA server and selects the target galaxies from the sample (those having spectroscopic redshift between 0.03 and 0.1). It then creates a tabular gridded datacube with values associated to potential neighbours. These values are calculates for each point of a 3D space defined by the axes: *magnitude in r band*, *photometric redshift* and *sigma level of detection*. The input default values to build the parameterised datacube are:

- 14.5<mr<22.5 - step 0.5
- 0 <z<0.11 - step 0.01
- 0.1<sigma<3.2 - step 0.2

Auxiliary function libraries and scripts are also copied in the local file system, and the PYTHONPATH environmental variable is set to a value provided by the user as the Working Path of the digital experiment.

Other input values provided by the user are the database connection settings: hostname, login and password.

2.6.2 Sample Selection. Environment

This workflow takes as input the path of the tabular **.pckl* Python pickle dataset created in the previous workflow, as well as the database connection settings, and several criteria on how to filter the potential companions of the target galaxies. It provides a file with the SDSS identifiers of each target galaxy of the sample, and environmental estimators and radius where the 10^{th} companion has been found. The workflow looks for potential companions in radius ranging from 3Mpc to 11Mpc, with a step of 1Mpc. The user may modify different parameters at the input stage, as well as several limits and ranges needed in the filtering process. As in the previous workflow, other provided input values are the Working Path of the digital experiment and the database connection settings: hostname, login and password.

2.6.3 Luminosity Profiles. Preprocessing

This workflow compiles the information needed by subsequent workflows in a VOTable and edits metadata information comprised in image files so that they be analysed afterwards. The user needs to provide a configuration file with the absolute path of the folder where the RO is located in the local file system, and a set of paths relatives to it where the original image files may be found, and where the files generated by the workflows will be stored. Another input is a tabular ASCII file where each row contains information of an image: the coordinates, the image filename, SDSS metadata uniquely identifying the image, a flag indicating if the galaxy is barred or not, and the filenames for the PSFIELD and PSF files created in this workflow.

The workflow is composed by two nested workflows. One of them makes up the absolute paths of the files, extracts the band and the resolution of the images from the images headers and queries the *sdssdr7-field*[19] service to obtain the properties of the fields. It also downloads the PSFIELD file from the SDSS web archive, runs the READ_PSF program on the PSFIELD file in order to build the PSF file, and edits its header image to add needed information in a later analysis process. Finally, it stores those data as well as other provided by the user in VOTables. The second one compiles those VOTables into a single one together with the coordinates and the flag *barred* provided by the user, and it adds three more columns, which are calculated using the values of other columns.

Taverna runs out of memory when the workflow is preparing the sample of 282 images of this use case. This problem can be solved increasing the maximum of the 400MB default memory in the Taverna configuration file, or changing the preferences so that it does not keep the provenance in memory. In order to avoid changing the default settings of Taverna, the initial table with 282 rows may be also split in four parts, the workflow executed four times, and concatenate those four VOTables into a single one with a short shim workflow.

[19] http://vo.astronet.ru/sai_cas/conesearch?cat=sdssdr7&tab=field

2.6.4 Luminosity Profiles. SExtractor

The main idea of this workflow consists of using SExtractor to estimate a set of parameters that are needed in later stages of the experiment. The parameters to be inferred are initially declared in a configuration file located in the user file system. Other configuration files are produced at run time during the execution of the workflow.

The workflow receives as inputs a VOTable that encapsulates the information needed for its execution, a template to create SExtractor configuration files and vocabulary for them, as well as several string values designating the column in the VOTable that provide filenames for images, configuration files and result files.

Although the configuration files of SExtractor have a defined structure, the values of its parameters depend on the galaxy. Therefore, a template pattern to create the files is built and then key values in the template are replaced by values from the VOTable.

The workflow is composed of four nested workflows that are sequentially executed:

- Creation of configuration files: It fills the template for each galaxy present in the input VOTable and it checks that the terms used in the template and in the VOTable match the terms in the vocabulary.
- SExtractor execution: It runs SExtractor on each image/galaxy declared in the VOTable.
- Cat SExtractor results: SExtractor results are organized in different files. This workflow gathers all the information into a single VOTable that is joined to the original one.
- Addition of new information: It uses SExtractor results to insert new calculated columns in the VOTable. The resulting VOTable is the final result of the main workflow.

2.6.5 Luminosity Profiles. Galfit

This workflow uses GALFIT to fit Sérsic and exponential-disk functions to the brightness distribution of the galaxies in the images, in order to build a model of each galaxy. The execution of GALFIT requires a configuration file with information related to the image, the observation and the adjusting functions. It returns a text file with new values for the adjusted parameters and a multi-layer FITS file composed of 2-D images (*observed galaxy, model and residuals*).

The main inputs of this workflow are a set of template files and vocabulary for them, partition criteria to process barred and non-barred galaxies in a different way, and a VOTable that contains data and filenames. It also requires the specification of the VOTable column names that provide the information on where to store GALFIT results in the file system. The output consists of a VOTable as well as the standard products and intermediate files delivered by GALFIT executions.

Depending on the image and the configuration parameters, GALFIT may fail or crash. The workflow manages these errors by filling the VOTable with error values (-99 and -999) in order to make explicit in the VOTable that GALFIT failed.

The workflow is composed of six nested workflows that are executed sequentially:

- Creation of GALFIT configuration files: It fills the template for each galaxy found in the input VOTable and checks that the terms used in the template and in the VOTable match those in the vocabulary. The template is defined to make a good estimation of the centre.
- GALFIT execution: It runs GALFIT for each galaxy found in the VOTable.
- Addition of new information: It retrieves GALFIT results in order to create a VOTable that is joined to the original one.
- Creation of GALFIT configuration files for a second iteration: It fills a new template for each galaxy. The template is different depending on whether the galaxy has a bar (partition criteria). The template is defined to make a fine grain fitting by using the results of the first execution.
- GALFIT execution: It runs GALFIT using the second configuration files.
- Addition of new information: It retrieves GALFIT results in order to create a VOTable that is joined to the previous one and it adds new calculated columns that are needed in later stages of the experiment.

2.6.6 Luminosity Profiles. Ellipse

This workflow calculates the set of ellipses that better match the shape of a galaxy at different radii by performing two iterations. In the first one the centre of the galaxy and the outer ellipse are approximated since there might be bars that affect the inner ellipses. In the second one, a fine grain fitting is performed and output files with the data defining the ellipses are produced.

ELLIPSE execution depends on IRAF. In particular, the workflow uses *pyraf* as an intermediate python package to call IRAF functionalities. IRAF requires setting up an IRAF folder, where tasks and procedures have to be called from this folder. For this reason, IRAF path is one of the inputs of this workflow. The remaining inputs are a VOTable, two templates, a vocabulary and several column names to point out the VOTable columns that provide the required information.

Depending on the image and the SExtractor results ELLIPSE may fail or ask for new centres of the ellipses. The workflow detects this situation if ELLIPSE result files are not created and then it fills the VOTable with error values (-999) that allow filtering the table.

The workflow consists of the following steps:

- Filter VOTable: Although this workflow only needs information from SExtractor execution, the RO protocol is designed to add workflow results incrementally into the VOTables. For this reason, the VOTable used as input in this workflow is the one coming from *galfit* workflow, removing the rows (galaxies) that failed during GALFIT execution, which will improve the performance.
- Create ELLIPSE scripts: It uses the template pattern to create a *pyraf* script to run the first ELLIPSE iteration.
- Run first ELLIPSE iteration. It executes the previous *pyraf* scripts.
- Fill VOTable with ELLIPSE results: It retrieves information from ELLIPSE results and creates a VOTable, add new calculated columns, join the VOTable to the original one and filter the rows where ELLIPSE has failed.

- Create second new ELLIPSE scripts to obtain more refined results.
- Run second ELLIPSE iteration.
- Get VOTable without errors: It detects the galaxies where second iteration failed and returns a VOTable without these galaxies.

Although several intermediate files are created during the execution of this workflow, the final outcome is the last VOTable as well as the ELLIPSE data files produced in the second iteration.

2.6.7 Luminosity Profiles. Plots

This workflow produces plots of the luminosity profile for each galaxy found in the input VOTable. The generated plots are based on GALFIT and ELLIPSE results. In particular, the plots contain:

- Modelled bulge luminosity profile
- Modelled disk luminosity profile
- Modelled bar luminosity profile (if the galaxy has a bar)
- Combined modelled luminosity profile
- Observed galaxy luminosity profile (including error bars)
- Residual values of the subtraction of the profiles obtained from the model image and the observed one

Some of these luminosity profiles can be found in Appendix D.

2.7 Research Object management

We call RO-ification to the process of packing all the elements involved in a digital experiment into an integrated artefact encoding the experimental protocol, the related process and data, software dependencies on the execution environment, attribution and authoring, as well as the provenance of intermediate and final results. The only technology that we dispose in order to expose in a straightforward visual way the provenance and relationships among all RO components, their roles and their types, is the file system. The RO portal provides some incipient functionality to display these links both as a folder structure and as a prototype diagram. We propose a generic RO tree-folder structure that preserves workflow modularity for potential re-use and reproducibility of individually extracted workflows, and avoids the duplication of resources with the use of symbolic links, allowing RO content browsing from multiple perspectives.

In the following we expose the main notions behind the generic RO tree-folder structure. A detailed representation of Luminosity Profiles RO may be found in Appendix B. RO digital components are classified at a first level into bibliography, execution environment, data, processes and workflows. Bibliography, which is any kind of digitally published resource (PDF files, blogs, web pages, videos, posts, etc.), is split at a second level into used and produced. Information related with the execution environment is split into required configuration files, local software dependencies and web services used. We have given ASCII files the

names of the software dependencies and the *.dcr* extension to point out that these files are actually description text files containing information on how to set up the environment, special privileges, software versions, access control on web services, setting accounts, etc. We have split the data into two main folders, one of them containing all data involved in the experiment, and the other small subsets constituting example data needed to run the workflows for demonstration and reproducibility purposes. Data are then classified by their role in the RO: *user_input* (expected to be provided by the user)*, outcome* (produced in workflows, and most of them needed also as input for other worflows) *and results* (final products of the experiment). Finally, all these data are organized in different folders pertaining to their associated workflows. Given the high volume of some datasets (e.g. 6.8 GB of the postgreSQL database), we have provided ASCII files with the names of the datasets as filenames, and the *.lnk* extension, to point out that these files are actually description text files containing information on how to access the actual datasets that are published in the AMIGA server.

While *outcome* data are univocally related to the workflow that produce them, data provided by the user *(user_input)* may be used in one specific workflow or in several ones all along the RO protocol. That's the reason why *user_input* data are further split into *common* and *self* folders. Those datasets contained in the *self* folder are then classified into folders with the names of the workflows where they take part as input data. In the Luminosity Profiles RO, composed of five workflows, only two user input datasets are required: init_table.txt located in `/data/all/user_input/self/preprocessing` and a set of FITS files with 2D images of galaxies located in `/data/all/user_input/common/images`.

The *process* folder contains copies of the scripts (python and java beanshells) that are encoded inside the t2flow Taverna workflow files. We think that keeping a copy of the scripts in the file system provides a seamless inspection of the process code involved in the workflows, and improves the transparency. We also provide a folder *bin* where external binary files imported in Taverna for their on-the-fly execution, or references to external links, could be stored. In the Luminosity Profiles RO, we could have provided the *read_PSF* binary file as an external file to be downloaded and executed at run time, and hence avoid this local software dependency. We decided instead to document the local software dependency since this file may be different for both 32 bits and 64 bits execution environment. A symlink to `/config/ws` folder is created to account for the *web services* used as processes.

The *workflows* folder is split at a second level into main and nested workflows. The latter are small ones composing the first ones, and potentially being used in several main workflows. With the purpose of achieving re-use and reproducibility of individually extracted *main workflows*, we have organised their internal file system structure into folders accounting for their individual execution environment set-up, datasets and processes involved, as well as the provenance of the results. In order to avoid duplication of files in the RO structure, many of these folders and files are symlinks pointing to the same resources already present in other folders, which this is the case for the execution environment set-up, datasets and processes involved, as well as the final produced results. e.g. `/workflows/main/ellipse/data/subset/outcome/` points to folder /`data/subset/outcome/ellipse/`

Most of the workflow inputs in a RO are not provided by the user, but produced in previously executed workflows, which results in a duplicity of roles of datasets (input and output) belonging to the same RO. These issues have been solved providing symlinks inside the *income* folder pointing to output datasets produced in previous workflows, as well as differentiating machine generated input datasets from those actually provided by the user by locating them in different folders (*income* and *user_input*). Moreover, since not all outputs are used as inputs and not all inputs are machine generated, we have called these folders *income* and *outcome* so to clearly remark this fact.

Provenance of the workflow runs are stored inside *prov_runs* folder, differentiating runs executed with all the data or with only a small representative subsample. The t2flow Taverna workflow file together with a README.txt file, providing information on very specific issues, complete the list of resources foreseen for each of the workflow folders internal structure. We also propose a README.txt file at the first level of the RO folder structure, providing information related to the whole RO, as well as CONTENT.txt and LICENCE.txt files. The experimental protocol to follow in the whole execution of the RO could be detailed in the HOWTO.pdf file placed in the configuration folder.

We have also used the RO Manager[20] [9] developed in the Wf4Ever project to annotate the digital resources previously arranged following the generic RO tree-folder structure. Some of these annotations try to map the information expressed in the RO file system structure into semantic models developed in the project (wf4ever, wfdesc, wfprov), while others are needed because of the impossibility to express them in the file system hierarchy. Given the high number of resources to be annotated as well as the many different kind of semantic annotations taken into account, we developed a *shell script* to automate the process. All annotations and RO Manager commands are registered in this *shell script,* which allows to rebuild the RO at any moment.

Annotation mapping relationships and other information exposed in the tree-folder structure

- A document is a produced publication in bibliography
- A document is a referenced/used publication in bibliography
- All files in `/config/files` are configuration files
- All small files in `/config/soft` represent needed software dependencies
- All small files in `/config/ws` represent web services used
- All files in subfolders of `/data/all/user_input/` are user-provided required files
- All files in `/data/all/user_input/wfname` are user-provided required files for workflow *wfname*
- All files in of `/data/all/user_input/common` are user-provided required files shared by several workflows
- All files in subfolders of `/data/all/outcome/` are files produced in workflows
- All files in subfolders of `/data/all/outcome/wfname` are files produced by workflow *wfname*
- All files in subfolders of `/data/all/results` are actual final results of the RO

[20] https://github.com/wf4ever/ro-manager

- All files in `/data/subset` are sample data used to check reproducibility and repeatability of workflows
- All files in `/process/bin/` are binary files imported in Taverna workflows
- All files in `/process/scripts/` are scripts inserted into Taverna workflows
- All t2flow files inside `/workflows/main` are the actual workflows used in the RO
- All t2flow files inside `/workflows/nested` are small workflows inserted into one or several main workflows
- All files and symlinks inside `/workflows/main/wfname` are related to workflow *wfname*
- All files in subfolders of `/workflows/main/*/income/` files are input files
- All files in subfolders of `/workflows/main/*/income/wfname2` are output files of workflow *wfname1*
- All files in subfolders of `/workflows/main/*/prov_runs` are provenance executions of a workflow
- Some of the products are plots, other are tabular data, other binary FITS 2D images

Annotations accounting for information not exposed in the tree-folder structure

- Author of an annotation
- Author and co-authors of a workflow; reference link to a re-used workflow and its author
- Who has performed the execution of a workflow leading to the results provided in the RO
- Computing execution environment of the RO and local software dependencies
- Special access requirements to web services
- Datasets provider: person, webpage, survey, data release, etc.
- How much time does it take to run a workflow using the full data and the provided subsample
- The number of elements of the sample dataset where one workflow and/or RO iterates
- Previous and subsequent workflows to be executed, as described in the experimental protocol
- Research institution, country, and scientific domain of the RO
- The actual size of the RO and/or a folder
- The version of a workflow

3. Results

3.1 Characterization of the environment for a sample of galaxies

Two environmental parameters were derived using the Sample Selection RO. The first parameter traces the density of local companions and the second one traces the tidal forces exerted by the companions. The estimation was obtained for a sample of ~300000 galaxies with redshift between 0.03 and 0.1 covered by the SDSS spectroscopic survey.

The relation of these parameters with some properties of the galaxies was studied. A correlation between the density and tidal parameters and the presence of an Active Galactic Nuclei (AGN) was found. Galaxies located in denser environments (high value of the density parameter) present less prevalence of those AGN selected based on the properties of the optical spectrum, but a higher prevalence of radio jets in their cores (radio AGN). On the other hand, a higher value of the tidal estimator is correlated with a higher prevalence of radio AGN and optically selected AGN.

Our interpretation of the results is as follows. Galaxies in denser environments do not have the supply of cold gas that is required to power optically selected AGN. However, warm gas can power radio AGN in these dense environments. On the other hand, all types of close interactions with companions (traced by a higher value of the tidal estimator) produce an enhancement on the prevalence of all types of AGN. The interaction with companion galaxies can fuel the gas to the centre of galaxies where is accreted by the central supermassive black hole producing both radio and optical AGN.

3.2 Luminosity profiles

The Luminosity Profiles RO developed in order to obtain luminosity profiles of galaxies was applied to a subsample of 90 isolated galaxies present in the AMIGA Catalogue [7] and using three different photometric bands from the SDSS: g, r, and i-bands. The 2-D models of the bulge component of galaxies obtained with GALFIT confirm that most of our galaxies have pseudobulges (according to their Sérsic index), which, according to the bibliography is indicative of secular evolution processes. The same result is obtained in the three bands, making hence the result more robust.

The analysis of the luminosity profiles reveals that some of the galaxies present a truncation in the ligh for their outer parts, meaning a faster decrease in the disk luminosity profile. We found also anti-truncations and normal profiles for other galaxies in our sample of isolated galaxies. This rejects the interaction with other galaxies as the mechanism responsible for the formation of some of these profiles, has has been suggested in the bibliography. Some of these luminosity profiles may be found in Appendix D.

4. Discussion

4.1 Impact

The two different approaches followed in the design of the workflows and ROs of this GE have provided different feedback on the potential impact on the working methodology of astronomers. As explained before, one of them migrated existing scripts that provide automation of tasks for a very big sample of datasets, while the other migrated actions performed on local interactive software in order to analyse a smaller sample of data.

Automation of tasks is a pressing concern that has been successfully solved with scripting in different program languages and environments, depending on the specific astronomical domain of research. The added value arising from the migration of existing scripts into workflows is not the automation of the process, but the improvement in the transparency of the experimental protocol. This allows the astronomer to precisely know how to execute the experiment, which datasets are needed and how to set-up the execution environment. This knowledge is hidden in the scripts, preventing the reproducibility of the experiment and hampering the replicability of digital science.

Storing digital recipes as data generators instead of final data products is a must in the upcoming context of Big Data, as well as organisation and structuring of information. The *trial and error* experimental methodology results in redundant needless files that the scientist needs to line up and classify in order to save time in understanding and reproducing semi-manual tasks.

The RO tree-folder structure is an attempt to expose the scientific experiment in a structured way, easily understood while exploring and browsing its content. It allows grouping a list of resources into bigger components (folders) providing a comprehensive view of the overall processes and data types, even if transparency is partially lost when packing but totally retrieved when unpacking.

It has been proven to be particularly arduous to achieve fully transparent workflows when migrating existing scripts. As a matter of fact, we can be quite confident that most of these workflows will still contain small bits of code hidden in *beanshells* as black boxes. Since a black box cannot be broken into parts, the re-use of these workflows entails the danger of spreading false assumptions and bad methods.

Migration of the existing experimental protocol into a more automated digital flow has made us notice the potential impact of workflows used as *living tutorials.* Scientists may visualize the actions performed by the workflows as they progress in their executions, allowing them to practice self-learning by the example, which expedites training and avoids reinvention. We are convinced that digital libraries of workflows and research objects will boost the use of the existing rich and underused infrastructure of data in Astronomy, and Virtual Observatory archives in particular, since they will provide these missing *living tutorials* on how to use them. Big storage and computing distributed infrastructures (HPC, Grid, Cloud) will be also exploited due to easing the application of the developed methodology to larger samples requiring higher computational power. Astronomers in the AMIGA group are already benefiting of some of the workflows developed in this Golden Exemplar in order to re-use them as templates for similar experiments (e.g. extraction of luminosity profiles

in radioastronomical images). The RO folder structure may also help to achieve templates for RO, where re-use makes RO building easier than starting from scratch.

A caveat here is that researchers demand software engineers to develop their workflows and *workflowcentric* ROs. Solutions for this will have to be discussed in the last year of the project in order to evaluate how this pilot project can impact a wide community.

Existing studies have shown that the top barrier for the scientists to publish their results in a reproducible way is the time required for creating documentation [6] These practices are not only laborious and time consuming, but what is worse, they are not properly rewarded. The only obvious incentives that we find to produce a fully documented and annotated RO are: later re-use (more efficient) by the creator, sharing within the research group/collaborators, training of e.g. new PhD students. Making those methods public has some handicaps or barriers: a) editorials do not ask for them, what is more astonishing, often do not even ask for the used data to be made available, b) lack of a proper citation methods imply sadly a risk of plagiarism, c) and directly related with the previous one, why helping the competitors in absence of rewards?

Other studies have shown that citation rates are higher for those articles where the scientists spent some time providing links to digitally published data [1] [2] A thoroughly annotated RO will greatly boost the visibility of the scientific research. This is where Wf4ever project can help: considering annotations as part of the weights in algorithms for recommendations, which in turn will raise the citation rates. Visibility is one of the most coveted goods in science, and it could be used as a highly significant incentive for providing well-documented digital experiments. The possibility to add external URL resources as “used bibliography” to the RO improves the visibility of any digital resource available on the Web (blogs, videos, posts, web pages, unpublished documents, etc.) Another contribution of Wf4ever will be scientist-friendly tool to provide annotations, as we will comment below.

4.2 RO Building and annotations

Most of the effort spent in the RO building process has been devoted to structuring RO content in the proposed RO tree-folder structure. We believe that some tooling would highly reduce the time spent in this process, since many of the files, folders and symlinks could be automatically generated from a predefined generic folder structure or RO template. Moreover, README and HOWTO files could be pre-populated with annotations and metadata present in t2flow Taverna files, as well as annotations performed with the RO Manager. Assistive tooling with proposed default attribute-value pairs would greatly help in providing semantic annotations compliant with existing semantic models or developed in the Wf4Ever project. As a first step we have contributed to the conception of an incipient *pattern catalogue* that could be used to develop assistive annotations tooling. This *pattern catalogue* intends to cover the most needed and common annotations in RO and RO components.

Annotations related to the experimental protocol, more specifically, the timeline story in each of the steps that have been followed, are among the most valuable when seeking the reproducibility of the experiment. Semantic models still do not cover many of these annotations, which usually fall into a common drawer where the information is not organised (e.g. purpose of research, not well-solved issues,

assumptions made, hypothesis to prove, caveats to consider, etc.) The working environment in the development of the experiment is Taverna workbench, and it is there where the user writes down those descriptions. A functionality to import annotations and descriptions from t2flow Taverna files would avoid duplication of annotations and would highly reduce the time in the annotation process, even if Taverna workbench does not offer the possibility to provide structured fine grain annotations. In this context of duplicated annotations, it would also be highly desirable to work in an environment that centralises the management of annotations for elements that are copied by reference like nested workflows or imported external scripts in beanshells.

Because ROs are composed of several workflows, the scientist needs to know the protocol to follow in their procedural execution in order to produce successful meaningful results. This information is provided in the HOWTO.pdf file and we believe it really helps in the understanding the addition of a RO *flowchart* as a picture/schema exposing the steps that actually interconnect them.

We have provided feedback for improvement of the RO Manager tool. Among the most important requirements, there is the need for control access management and the inclusion of user sharing capabilities. It should also be possible to link annotations to users, as well as functionalities to edit and remove user-accessible annotations. The big number of resources composing the RO needs to annotate multiple files with the use of wildcards, regular expressions and patterns, as well as filtering their display. We believe that re-use of existing ROs is key in spreading this working methodology. A functionality to import ROs from zip/tar compressed files or from an accessible URI would highly boost their use.

We have found that publishing a RO composed of a high number of files takes too much time and is highly inefficient. We have grouped and compacted files into single zip/tar files as often as possible (datasets and Taverna exported provenance files), and we have externalised big sized datasets as compressed files provided by the AMIGA web server. Seamless synchronization between local copies of ROs and those published in the RODL Wf4Ever Sandbox should also be improved, the tool being smart enough to push only updated files when updating the RO in the DL.

Finally, we have realised the lack of specific vocabularies for the astronomical domain, as well as an upper layer of categories placed on top of semantic models, which will group the resources by their type or role in the experiment. Implementing conditional annotations could also be considered. This would provide internal consistency among annotations; some of them may not make any sense if others have been already provided.

4.3 Quality and completeness

Among the most relevant criteria for quality evaluation of workflows and ROs are those related with prevention of decay. Four main weaknesses have been identified when studying workflows decay [8] : volatile third-party resources, missing example data, missing execution environment and insufficient descriptions about workflows. We believe that the proposed generic RO tree-folder structure would help in checking the missing pieces, since it provides placeholders for example data, characterization of the execution environment, as well as README and HOWTO files to describe the workflows and experimental

protocol. Information on the sequencing order in the workflows execution, if mapped to models, could be really helpful when developing checking tools for liveliness, consistency and completeness. In the development of workflows we have adopted the strategy to implement errors and exceptions management in the workflow design from its conception, which minimises the risks of decay and improve the quality.

Completeness checking is needed not only to assess the quality of the RO and workflows, but also as a prerequisite to proceed for quality evaluation based on other criteria like repeatability, reproducibility, liveliness or consistency, since the evaluation process is only possible when a minimum number of resources is present in the RO. Completeness checking for annotations could be seen as the minimum needed annotations that have to be provided in order to achieve a fully consistent RO according to semantic models developed in the Wf4Ever project. Users do not feel the need to export the provenance files in very specific formats, neither to know if they have covered the whole list of models attributes in the process of annotations, since they do not have the tools to exploit and visualize this information.

Quality assessment and completeness checking could play the role of incentives for producing well-documented and completed ROs, enhancing visibility in the community for those having better scores, in a similar way to Google PageRank metrics.

4.4 Preservation and evolution

We have been forced to take very difficult decisions related with local software dependencies in the conception and design of workflows. We are aware that preservation is more reliable for autonomous self-sufficient packs. Our ROs and workflows are mostly based on local software dependencies, or in the best cases rely on external web services or data. This has been the price we had to pay in order to achieve the migration of experimental protocols and processes (software) that are familiar to astronomers. We could mitigate this issue importing external software binary files and enacting them at workflow run time. Taverna workbench offers the possibility to declare binary files and scripts “by reference”, which may be downloaded at the moment of execution. A complete externalisation of the processes would need a fully monitoring of these external resources, as it has been already suggested for the web services in Wf4Ever project forums. In this optic, the possibility to link published nested workflows in the Taverna workbench “by reference” and not “by value” would be welcome.

Preservation is deeply related with conservation tasks performed on the archived ROs when facing a potential reconstruction in order to bring a RO back to life. At this stage, provided annotations pertaining the recovering process play a crucial role. Information on authoring and credit attribution is useful to achieve long sought citation rates, but most important this entails responsibility. We consider of great relevance the possibility to register the tuples *user-annotation*, as well as the provenance indicators *“who, when and why”* for workflow runs, in order to know who to blame and asked for specific issues related his contribution. This practice should in principle modify the existing citation system, enabling credit attribution to specific parts of the experiment as well as different roles in the contribution.

The creation of the RO is done after the experimental process, just before final publication step. It can be considered as a needed step to clean data and pack methods in a tidy structured way. In principle,

there should not be evolution or versioning issues in the RO-ification process, but mainly because of the fact that different users may contribute to the process makes these concerns arise. We have identified the need of a mechanism to rebuild ROs from previously registered actions performed with the RO Manager in order to rebuild different versions of the proto-RO, which we have solved with the help of versioned shell scripts.

RO evolution after publishing is a matter we have not covered, nevertheless we have been providing our user feedback to these discussions inside the project forums. A non-straightforward point in evolution is *identity* of the RO after re-use: when should we consider a workflow or RO to be an upgrade of the same one, or a different entity?. ROs and workflows should keep the history of its ancestors when they are created from forking or re-used by the same author. The re-use by another user would imply the creation of a new entity, with no history associated to this author. We think that there is a history related to the "author-artefact" tuple, where the user is very interested in the evolution of e.g. integrity, stability, completeness, or quality, and a history related to its previous lives as different entities. A too restrictive definition of when an RO evolves to new version, forcing it to become a new RO implies a quick loose of its evolution indicators.

5. Conclusions

We have reached a stage in the Wf4Ever project where a digital experiment is no longer bundled into a set of several scientific workflows and their associated digital artefacts (datasets, process, provenance, metadata, etc.), but into a set of several ROs. The 2nd Astronomy Golden Exemplar is composed of two different ROs that have been conceived and designed with two different approaches: the migration to workflows of already existing scripts focusing on the automation of big datasets crunching, and the migration to workflows of well-known detailed experimental user protocols followed in astronomical research.

Both strategies have provided relevant insight in the process of RO-ification as well as in the decisions to take at the stage of conception and design of workflows and ROs. We have found that structuring of information, as opposed to processes automation, is a need that has to be solved. In this context, we have provided a generic RO tree-folder structure as a first attempt to expose in a straightforward visual way the roles and types of all the RO components, as well as their provenance and inter-relationships. We believe that RO tree-folder structure would help in developing tools for completeness checking, since it provides placeholders for those elements identified as the main responsible of decay when they are missing.

Transparency of the scientific methodology exposed in workflows is another relevant added value to take into account when we decide to migrate existing solutions to workflows. Even if we failed in achieving fully transparent workflows when migrating existing scripts, we detect a real improvement in the transparency of the experimental protocol. i.e. how to run the experiment, which datasets are needed, how to set-up the execution environment. Adding to this fact the benefits arising from having a structured pack thoroughly characterised with appropriate metadata, workflow migration and subsequent RO-ification is already worthy of consideration. Nevertheless, we have detected that there is still information that cannot be covered with semantic annotations; users faced to the fact that they do not have the tools to explore them, neither the way to express them.

Finally, we would like to stress on re-use as one of the main trails to achieve incremental scientific development. Providing simple but representative workflows and ROs as templates improves re-use and self-learning by the example. Astronomers of the AMIGA group have understood the main processes involved in the workflows developed, which has triggered new ideas on how to re-use them to solve similar problems in their own research.

6. References

[1] E. A. Henneken, M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. Grant, D. Thompson, S. Murray. “Effect of E-printing on Citation Rates in Astronomy and Physics” Journal of Electronic Publishing, 9, 2. 2006

[2] E. A. Henneken, A. Accomazzi “Linking to Data - Effect on Citation Rates in Astronomy," ArXiv e-prints, arXiv:1111.3618 (2011)

[3] V. Karachentseva. “Catalogue of isolated galaxies” Comm. Spec. Ap. Obs., 8, 1, 1973.

[4] R. Palma, M. Werla, M. Nowak, M. Matela, P. Hołubowicz, S. Soiland-Reyes, J. Bhagat, D. De Roure, G. Klyne. “Wf4Ever Sandbox v1”. Technical report, D1.2v1, Wf4Ever Project, 2011.

[5] R. Palma, P. Hołubowicz, G. Klyne, A. Garrido. “Wf4Ever Sandbox – Phase II”. Technical report, D1.2v2, Wf4Ever Project, 2012.

[6] Stodden, V. “The scientific method in practice: reproducibility in the computational sciences” MIT Sloan Research Paper, 2010

[7] L. Verdes-Montenegro, J. Sulentic, G. Bergond, D. Espada, S. Leon, U. Lisenfeld, V. Martínez-Badenes, J. E. Ruiz, J. Sabater, and S. Verley, “A galaxy baseline : Multiwavelength study of a sample of the most isolated galaxies in the local universe,” in Galaxies and their Masks, 2010, pp. 1–6.

[8] J. Zhao, J. M. Gomez-Pérez, K. Belhajjame, G. Klyne, and E. García-Cuesta, A. Garrido, K. Hettne, M. Roos, D. De Roure, C. Goble “Why Workflows Break - Understanding and Combating Decay in Taverna Workflows.” Proceedings of the IEEE eScience Conference, 2012

[9] J. Zhao, G. Klyne, P. Holubowicz, K. Hettne, J.E. Ruiz, M. Roos, and J. M. Gomez-Pérez, D. De Roure and C. Goble “RO-Manager : A Tool for Creating and Manipulating Research Objects to Support Reproducibility and Reuse in Sciences,” Proceedings of the LISC2012 Conference, 2012.

Appendix A – Graphical representation of workflows

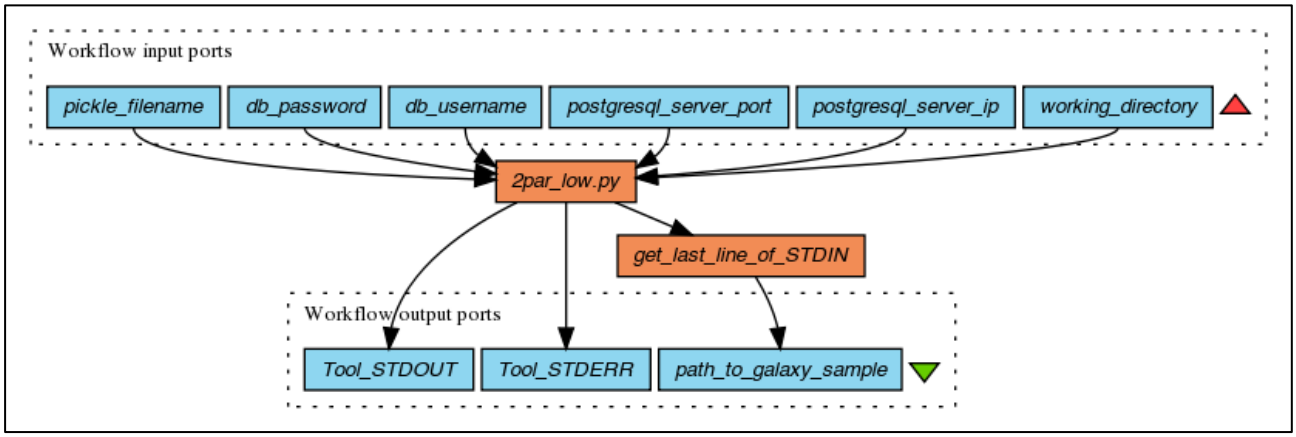


Figure 1: Initialize Workflow

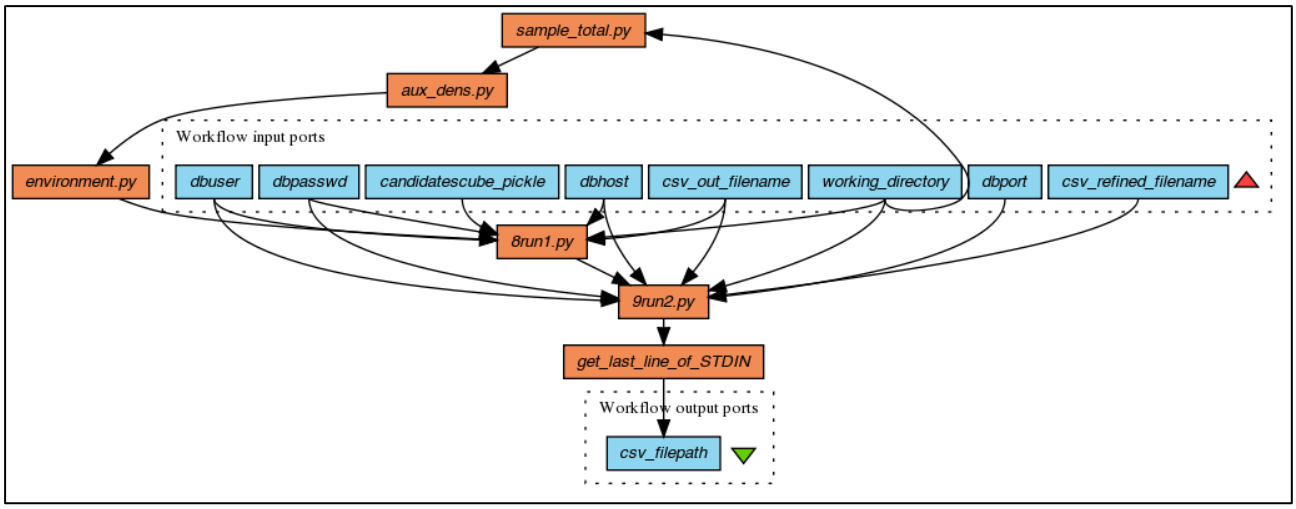


Figure 2: Environment Workflow

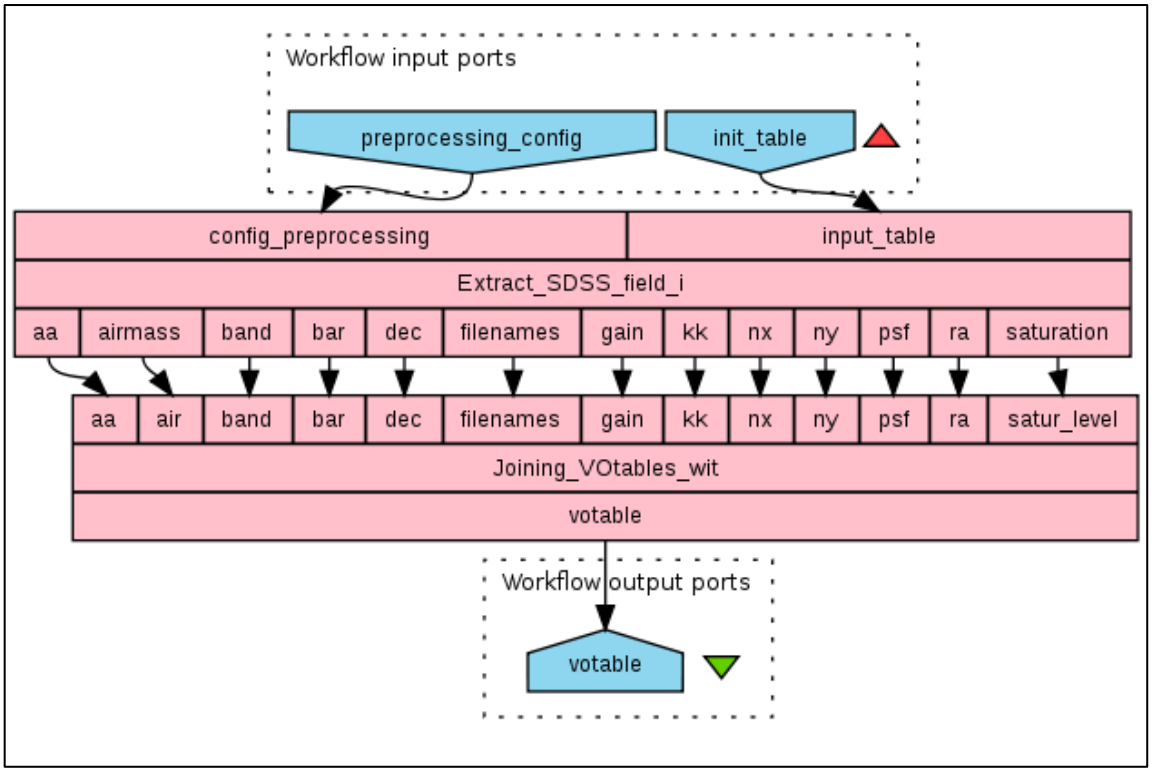


Figure 3: Preprocessing Workflow

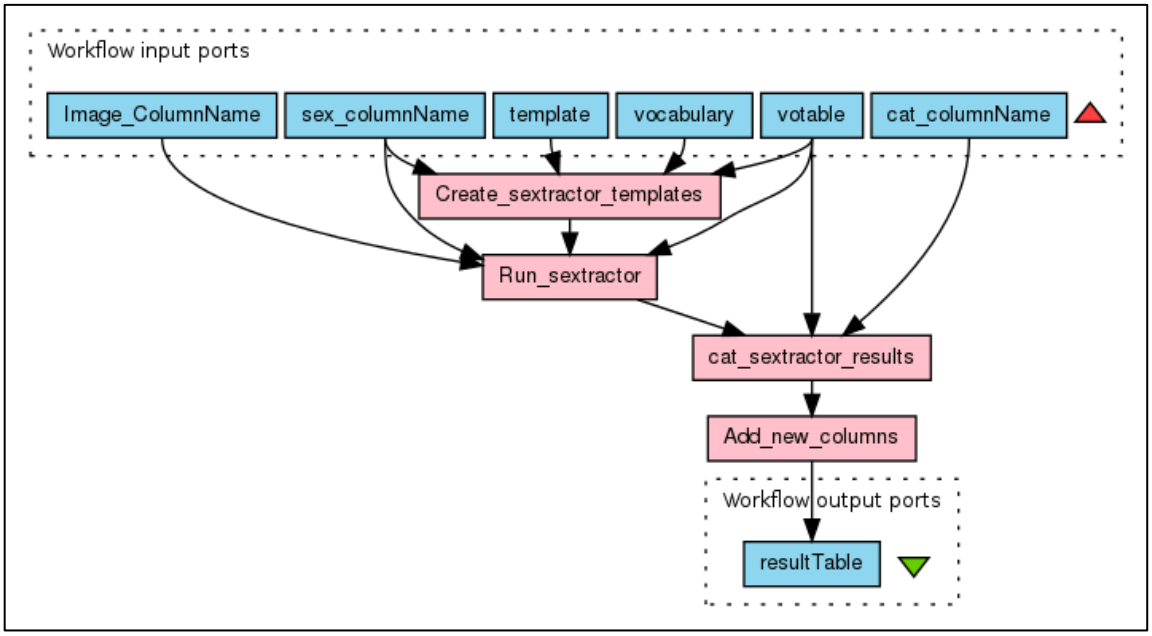


Figure 4: SExtractor Workflow

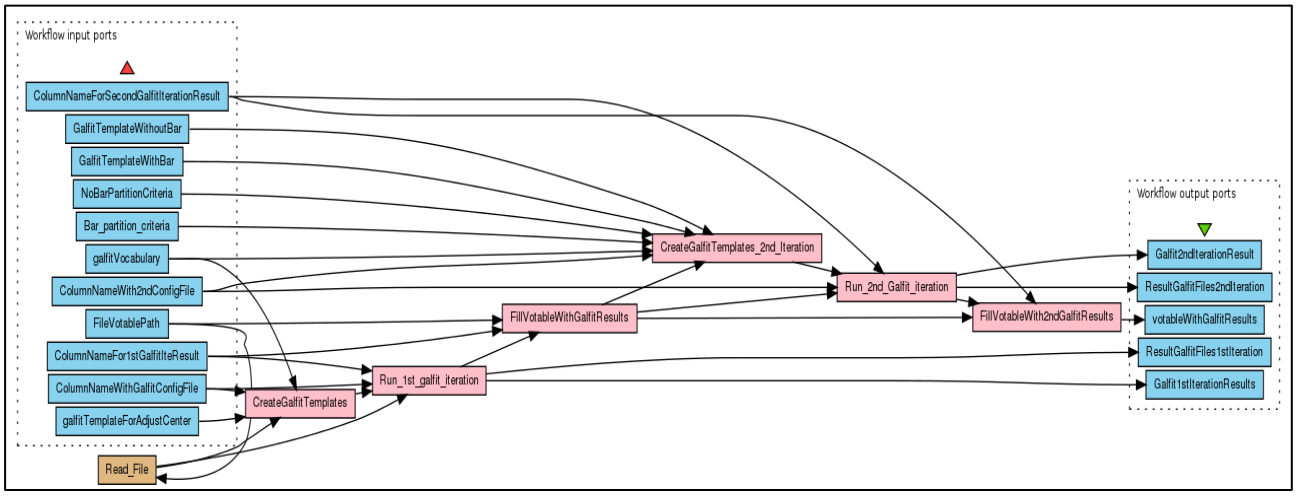


Figure 5: Galfit Workflow

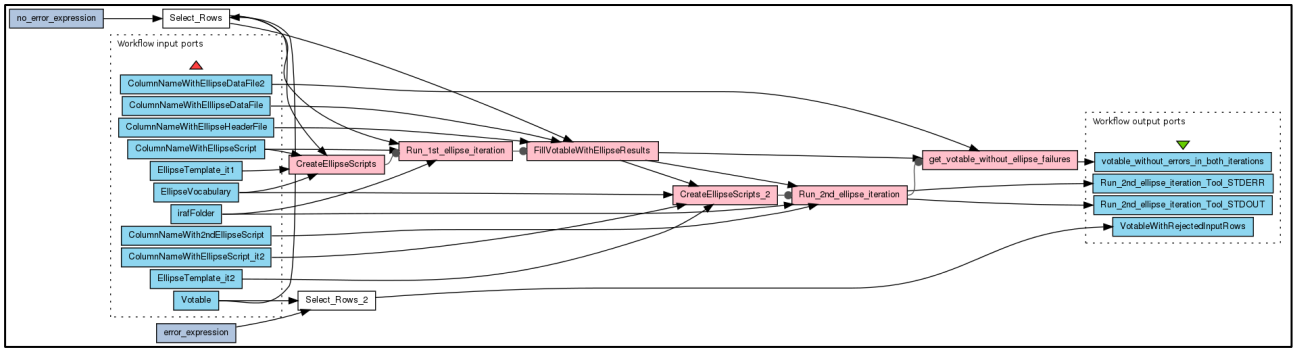


Figure 6: Ellipse Workflow

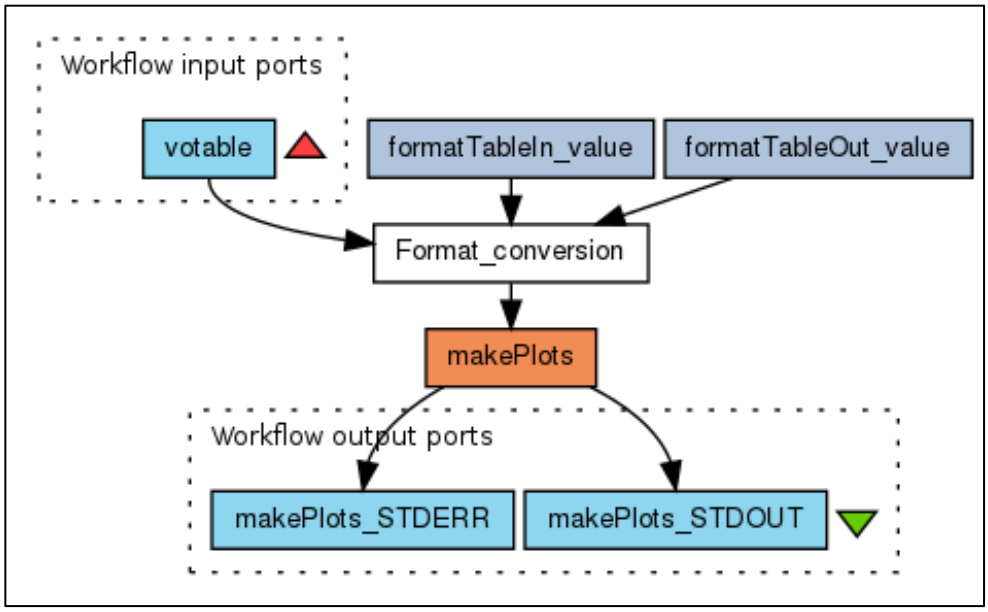
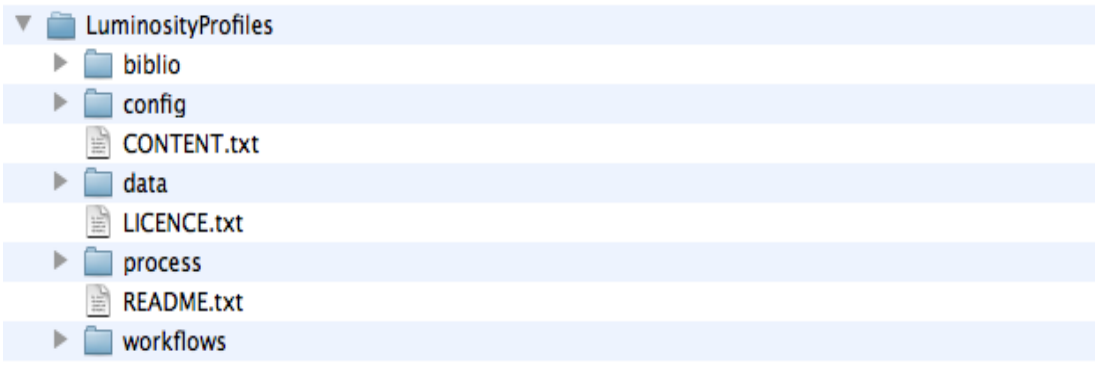


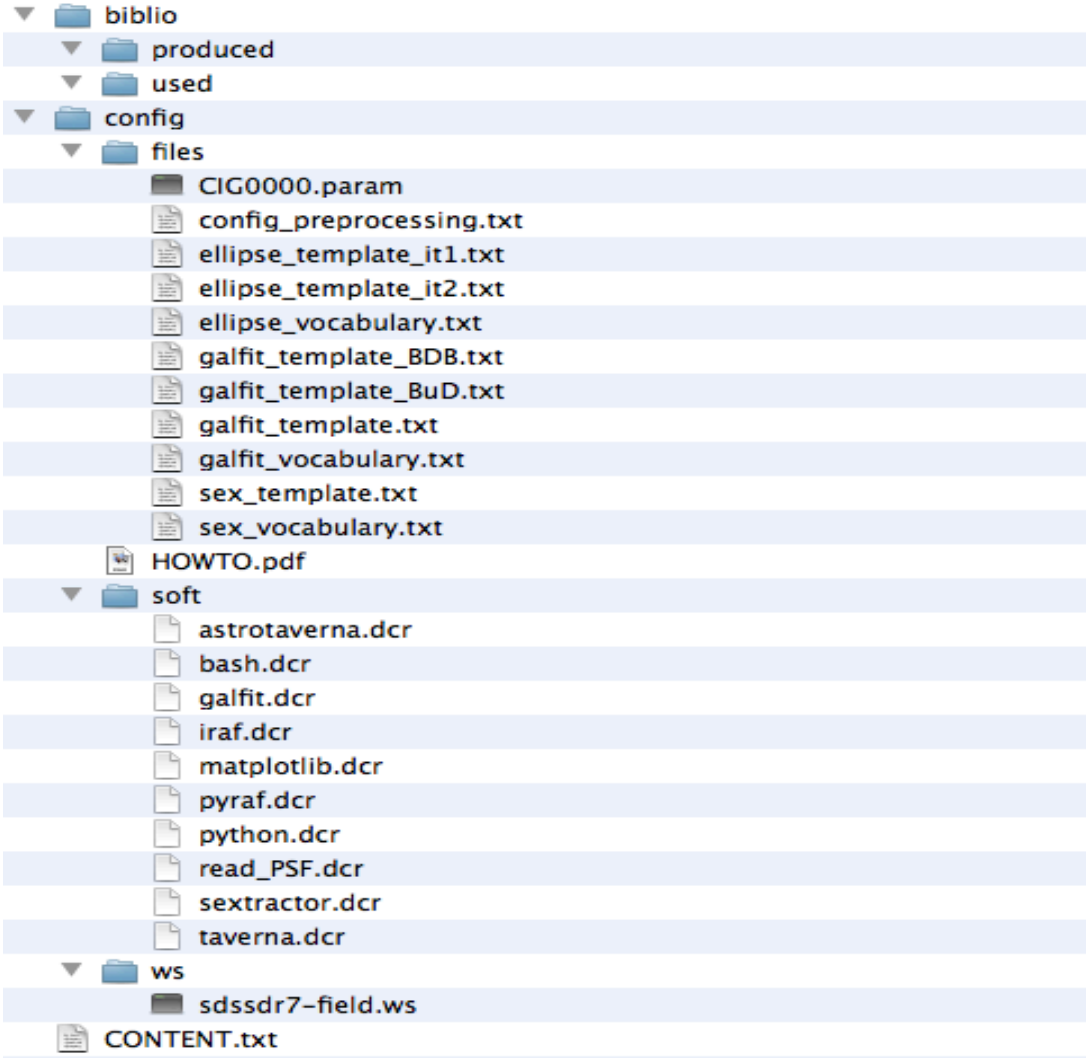
Figure 7: Plots Workflow

Appendix B – Luminosity Profiles RO structure and content

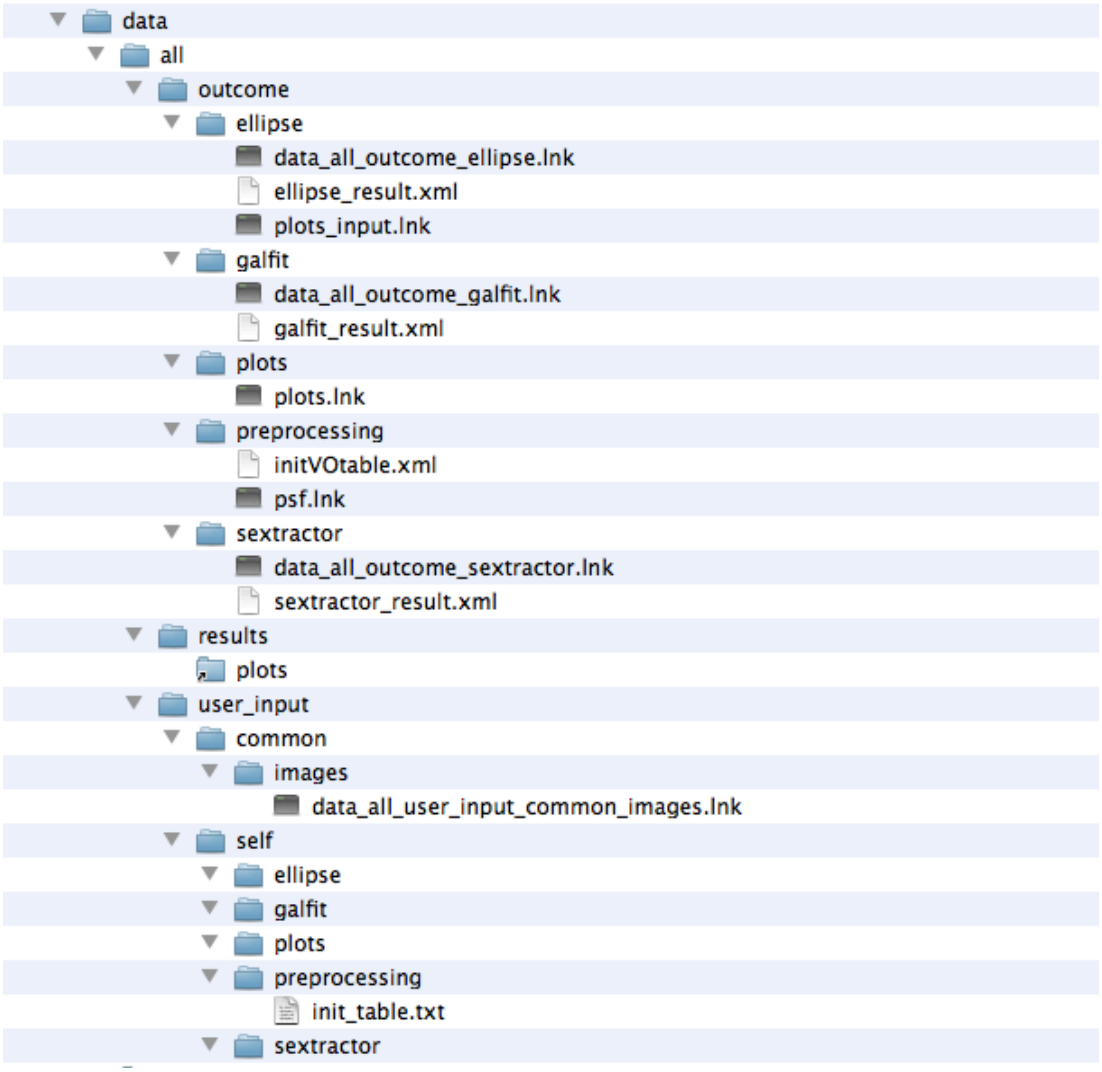
ROOT



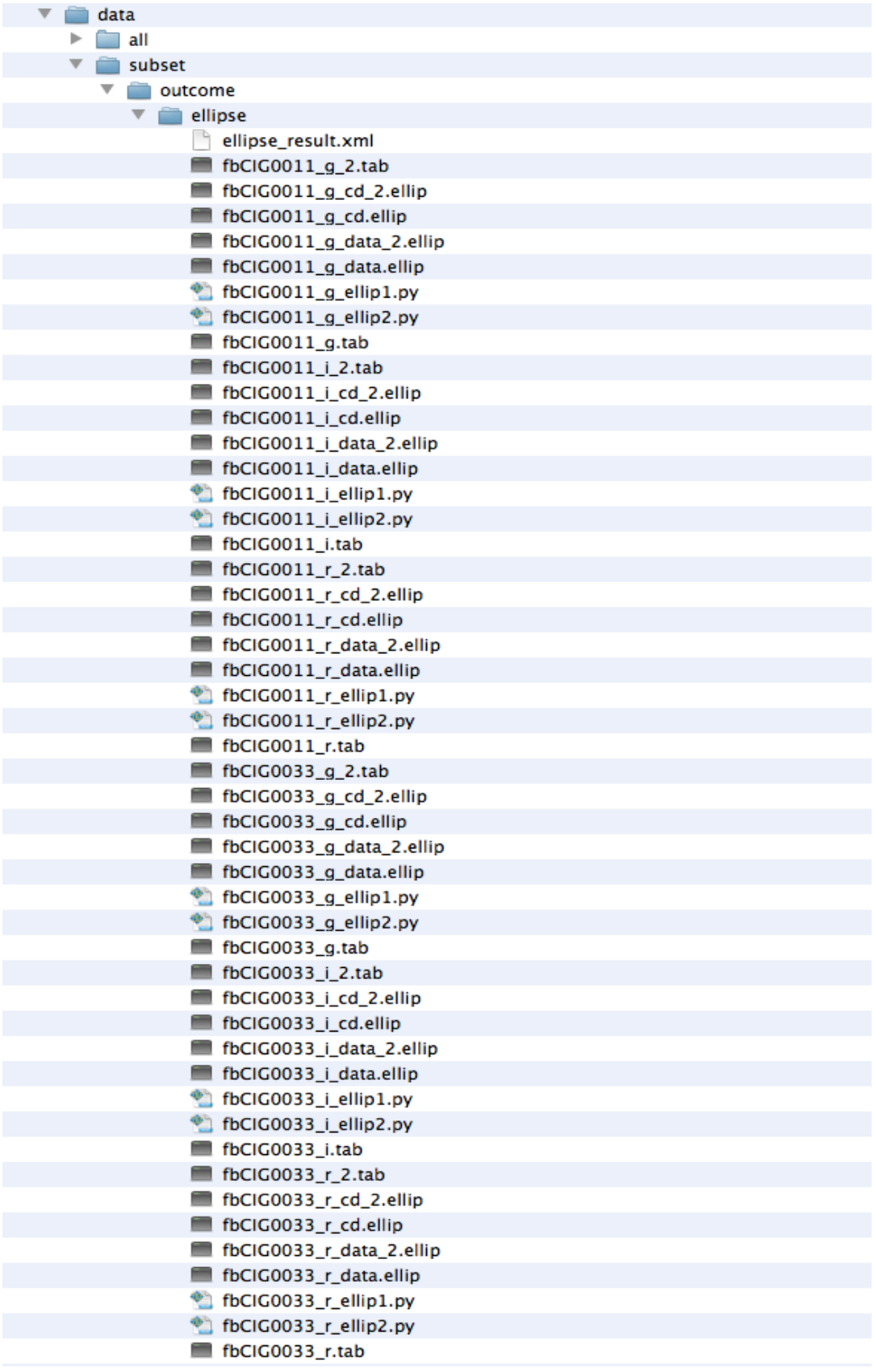
CONFIG

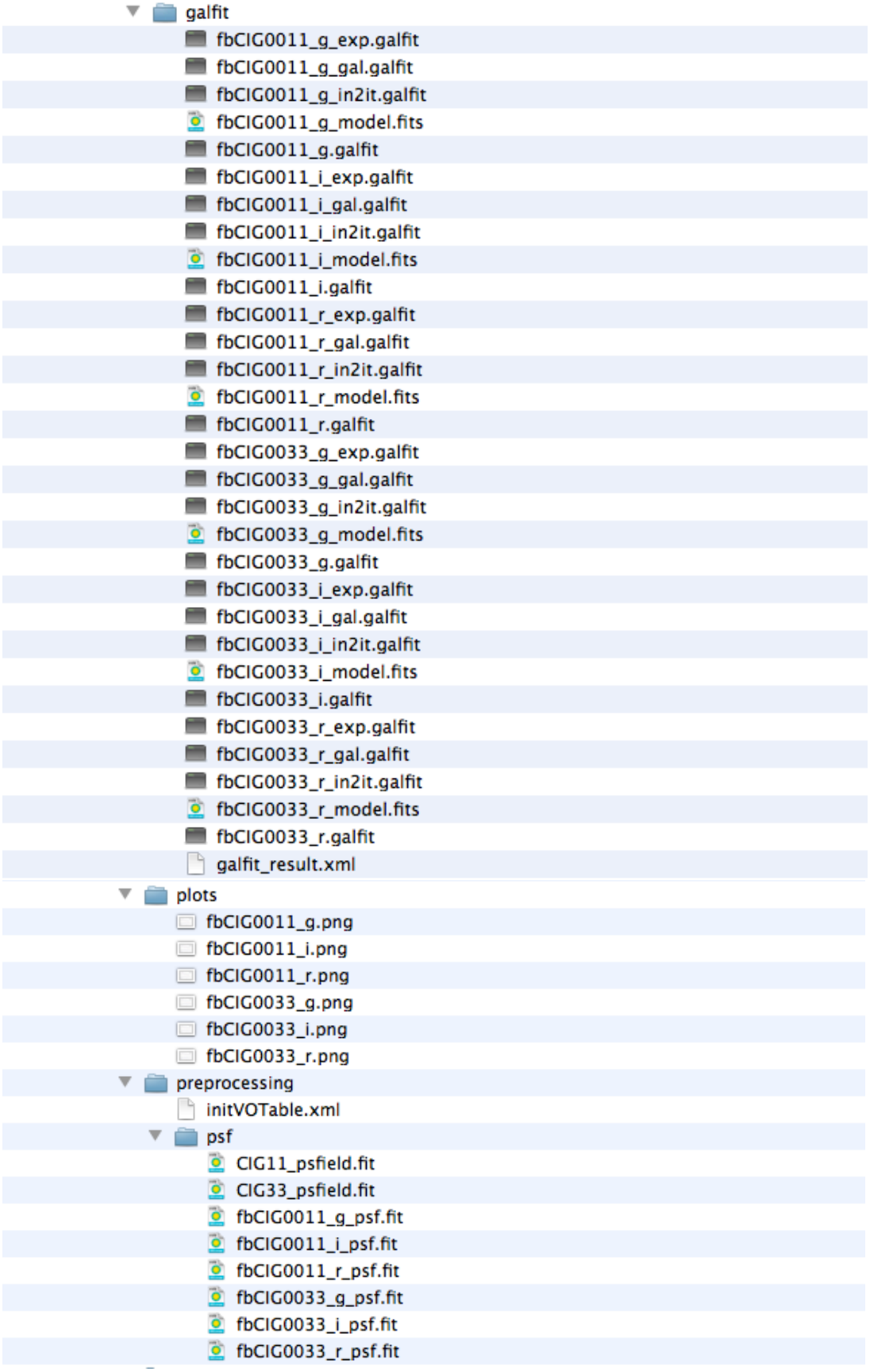


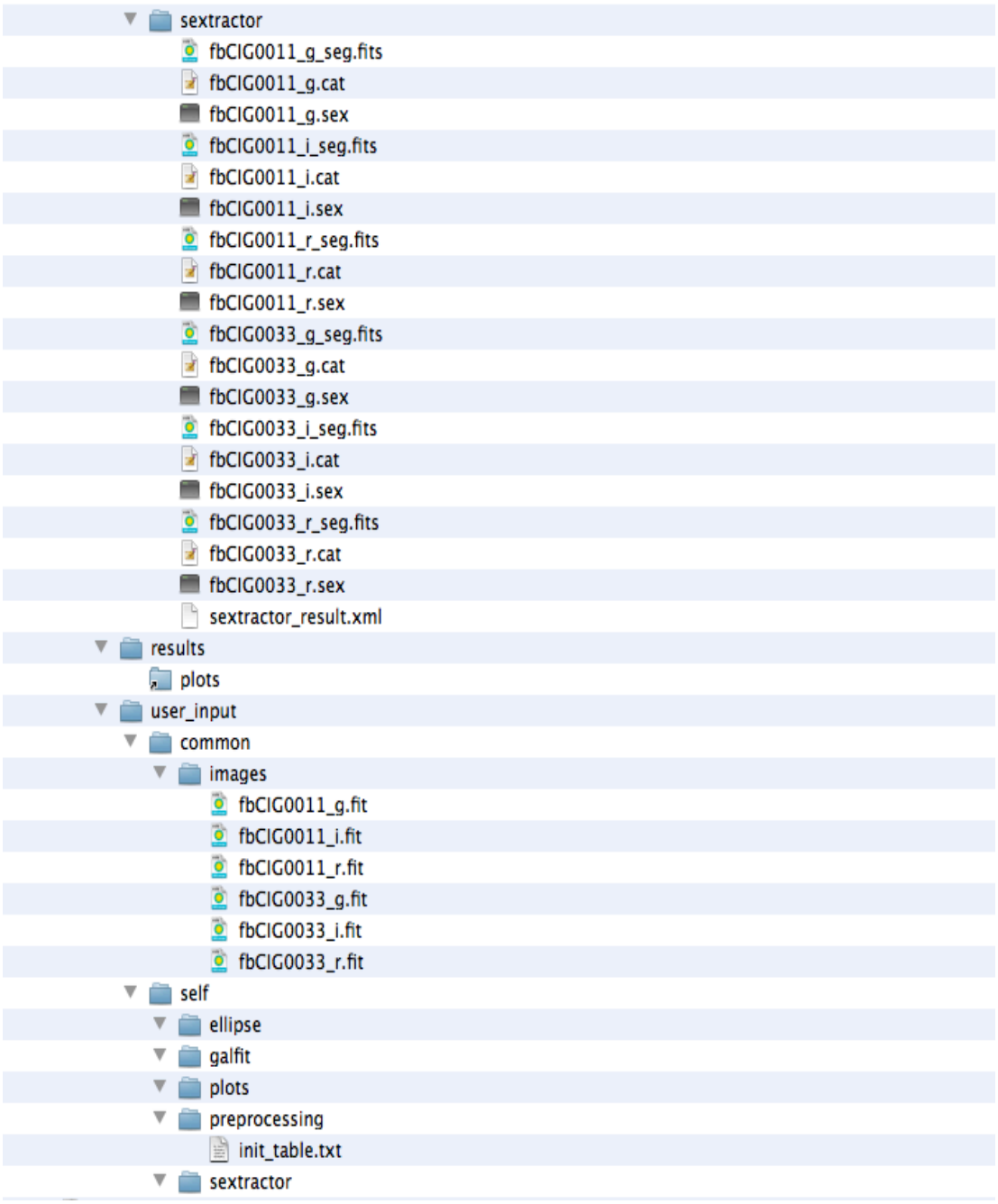
DATA


data
all
outcome
ellipse
data_all_outcome_ellipse.lnk
ellipse_result.xml
plots_input.lnk
galfit
data_all_outcome_galfit.lnk
galfit_result.xml
plots
plots.lnk
preprocessing
initVOtable.xml
psf.lnk
sextractor
data_all_outcome_sextractor.lnk
sextractor_result.xml
results
plots
user_input
common
images
data_all_user_input_common_images.lnk
self
ellipse
galfit
plots
preprocessing
init_table.txt
sextractor

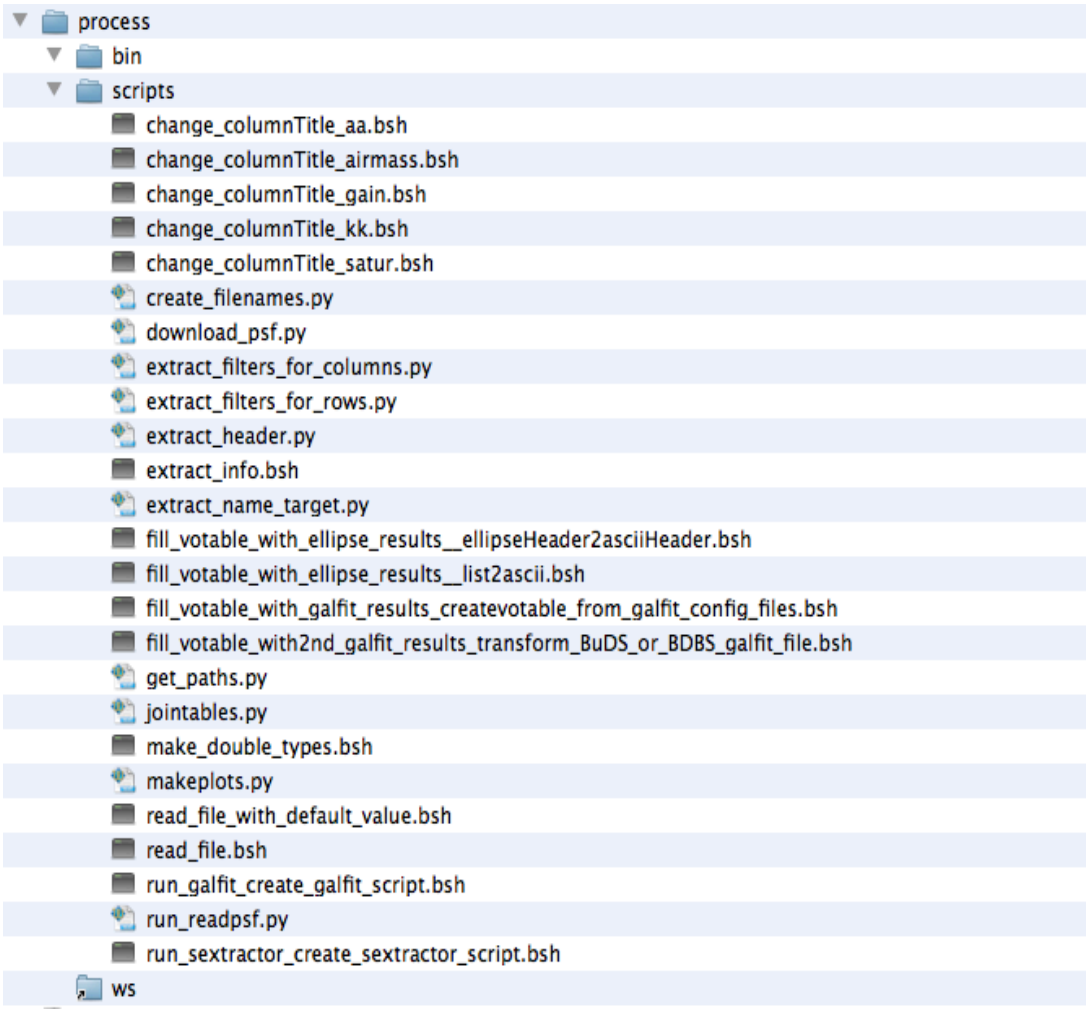
SUBSETS



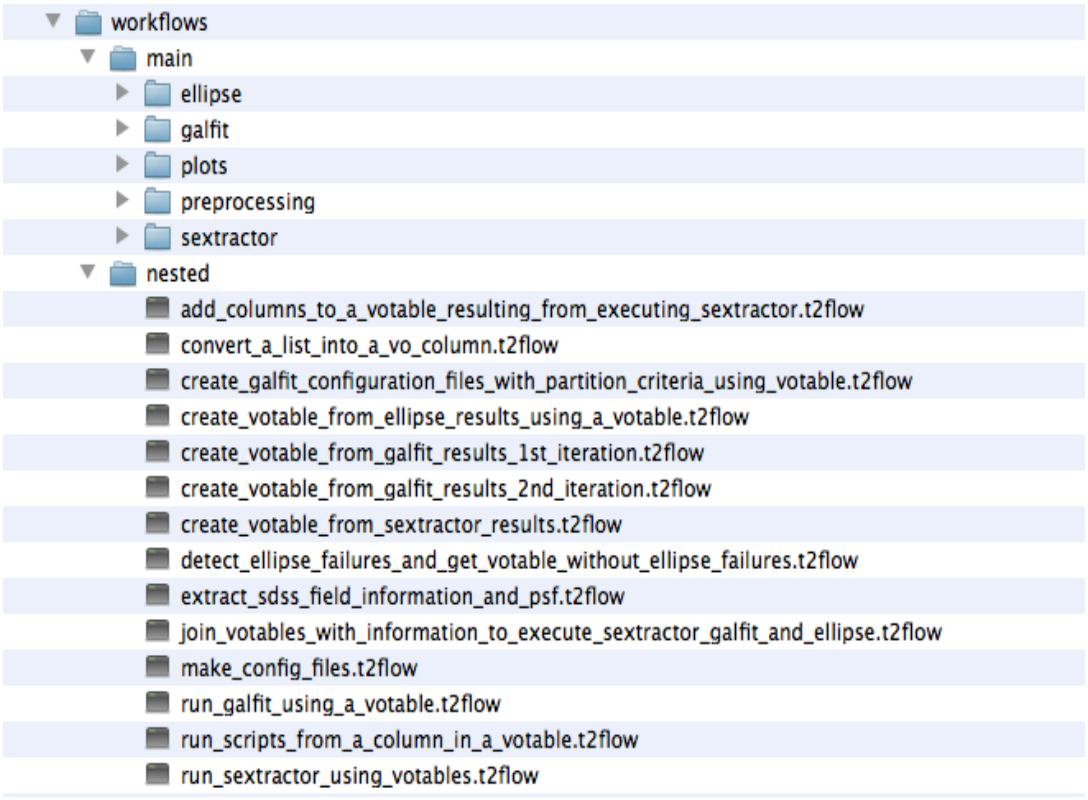

galfit
fbCIG0011_g_exp.galfit
fbCIG0011_g_gal.galfit
fbCIG0011_g_in2it.galfit
fbCIG0011_g_model.fits
fbCIG0011_g.galfit
fbCIG0011_i_exp.galfit
fbCIG0011_i_gal.galfit
fbCIG0011_i_in2it.galfit
fbCIG0011_i_model.fits
fbCIG0011_i.galfit
fbCIG0011_r_exp.galfit
fbCIG0011_r_gal.galfit
fbCIG0011_r_in2it.galfit
fbCIG0011_r_model.fits
fbCIG0011_r.galfit
fbCIG0033_g_exp.galfit
fbCIG0033_g_gal.galfit
fbCIG0033_g_in2it.galfit
fbCIG0033_g_model.fits
fbCIG0033_g.galfit
fbCIG0033_i_exp.galfit
fbCIG0033_i_gal.galfit
fbCIG0033_i_in2it.galfit
fbCIG0033_i_model.fits
fbCIG0033_i.galfit
fbCIG0033_r_exp.galfit
fbCIG0033_r_gal.galfit
fbCIG0033_r_in2it.galfit
fbCIG0033_r_model.fits
fbCIG0033_r.galfit
galfit_result.xml
plots
fbCIG0011_g.png
fbCIG0011_i.png
fbCIG0011_r.png
fbCIG0033_g.png
fbCIG0033_i.png
fbCIG0033_r.png
preprocessing
initVOTable.xml
psf
CIG11_psfield.fit
CIG33_psfield.fit
fbCIG0011_g_psf.fit
fbCIG0011_i_psf.fit
fbCIG0011_r_psf.fit
fbCIG0033_g_psf.fit
fbCIG0033_i_psf.fit
fbCIG0033_r_psf.fit



PROCESS

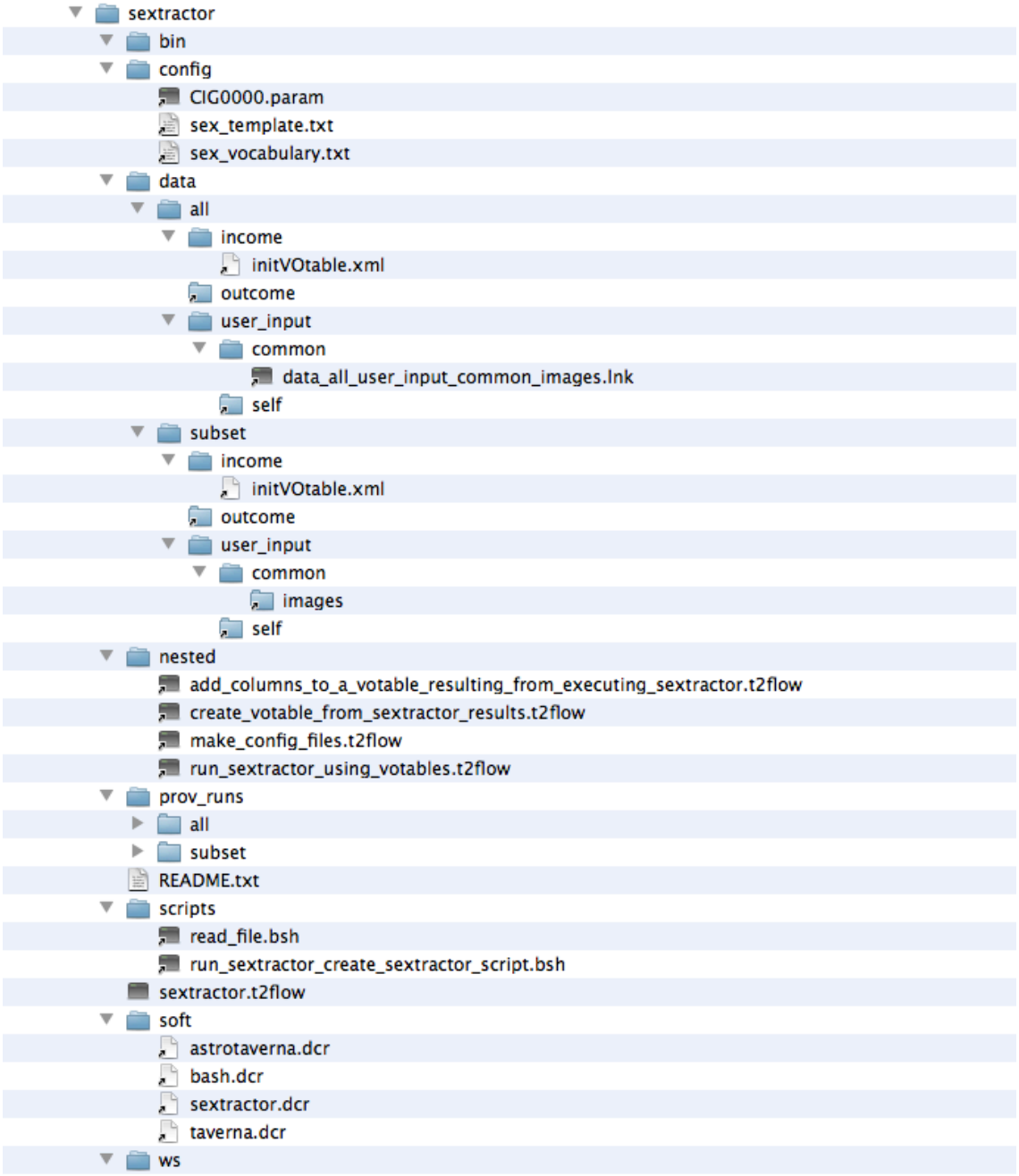

process
bin
scripts
change_columnTitle_aa.bsh
change_columnTitle_airmass.bsh
change_columnTitle_gain.bsh
change_columnTitle_kk.bsh
change_columnTitle_satur.bsh
create_filenames.py
download_psf.py
extract_filters_for_columns.py
extract_filters_for_rows.py
extract_header.py
extract_info.bsh
extract_name_target.py
fill_votable_with_ellipse_results__ellipseHeader2asciiHeader.bsh
fill_votable_with_ellipse_results__list2ascii.bsh
fill_votable_with_galfit_results_createvotable_from_galfit_config_files.bsh
fill_votable_with2nd_galfit_results_transform_BuDS_or_BDBS_galfit_file.bsh
get_paths.py
jointables.py
make_double_types.bsh
makeplots.py
read_file_with_default_value.bsh
read_file.bsh
run_galfit_create_galfit_script.bsh
run_readpsf.py
run_sextractor_create_sextractor_script.bsh
ws

WORKFLOWS



PREPROCESSING

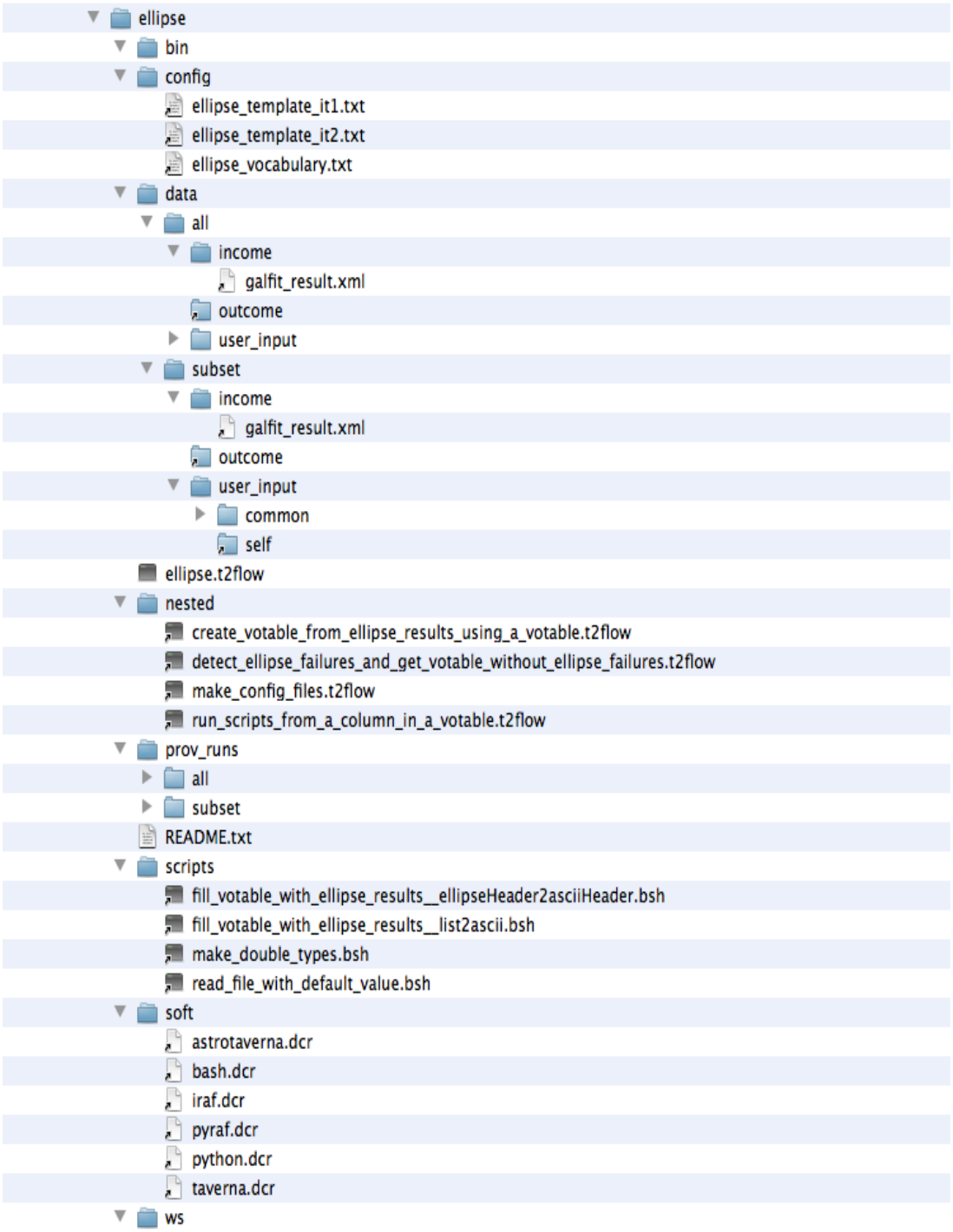
SEXTRACTOR



GALFIT


galfit
bin
config
galfit_template_BDB.txt
galfit_template_BuD.txt
galfit_template.txt
galfit_vocabulary.txt
data
all
income
psf.lnk
sextractor_result.xml
outcome
user_input
common
data_all_user_input_common_images.lnk
self
subset
income
psf
sextractor_result.xml
outcome
user_input
common
images
self
galfit.t2flow
nested
create_galfit_configuration_files_with_partition_criteria_using_votable.t2flow
create_votable_from_galfit_results_1st_iteration.t2flow
create_votable_from_galfit_results_2nd_iteration.t2flow
make_config_files.t2flow
run_galfit_using_a_votable.t2flow
prov_runs
all
subset
README.txt
scripts
fill_votable_with_galfit_results_createvotable_from_galfit_config_files.bsh
fill_votable_with2nd_galfit_results_transform_BuDS_or_BDBS_galfit_file.bsh
make_double_types.bsh
read_file.bsh
run_galfit_create_galfit_script.bsh
soft
astrotaverna.dcr
bash.dcr
galfit.dcr
taverna.dcr
ws

ELLIPSE



PLOTS


plots
bin
config
data
all
income
ellipse_result.xml
plots_input.lnk
outcome
user_input
common
self
subset
income
ellipse_result.xml
fbCIG0011_g_data_2.ellip
fbCIG0011_i_data_2.ellip
fbCIG0011_r_data_2.ellip
fbCIG0033_g_data_2.ellip
fbCIG0033_i_data_2.ellip
fbCIG0033_r_data_2.ellip
fbCIG0056_g_data_2.ellip
fbCIG0056_i_data_2.ellip
fbCIG0056_r_data_2.ellip
fbCIG0187_g_data_2.ellip
fbCIG0187_i_data_2.ellip
fbCIG0187_r_data_2.ellip
outcome
user_input
common
self
nested
plots.t2flow
prov_runs
all
subset
README.txt
scripts
makeplots.py
soft
matplotlib.dcr
python.dcr
taverna.dcr
ws

Appendix C – Results of Luminosity Profiles RO

Luminosity profiles obtained with ELLIPSE (black points). The solid lines are the model given by GALFIT for each component, and the total model (blue line). In the bottom panels we represented the difference between the observed data and the total model.

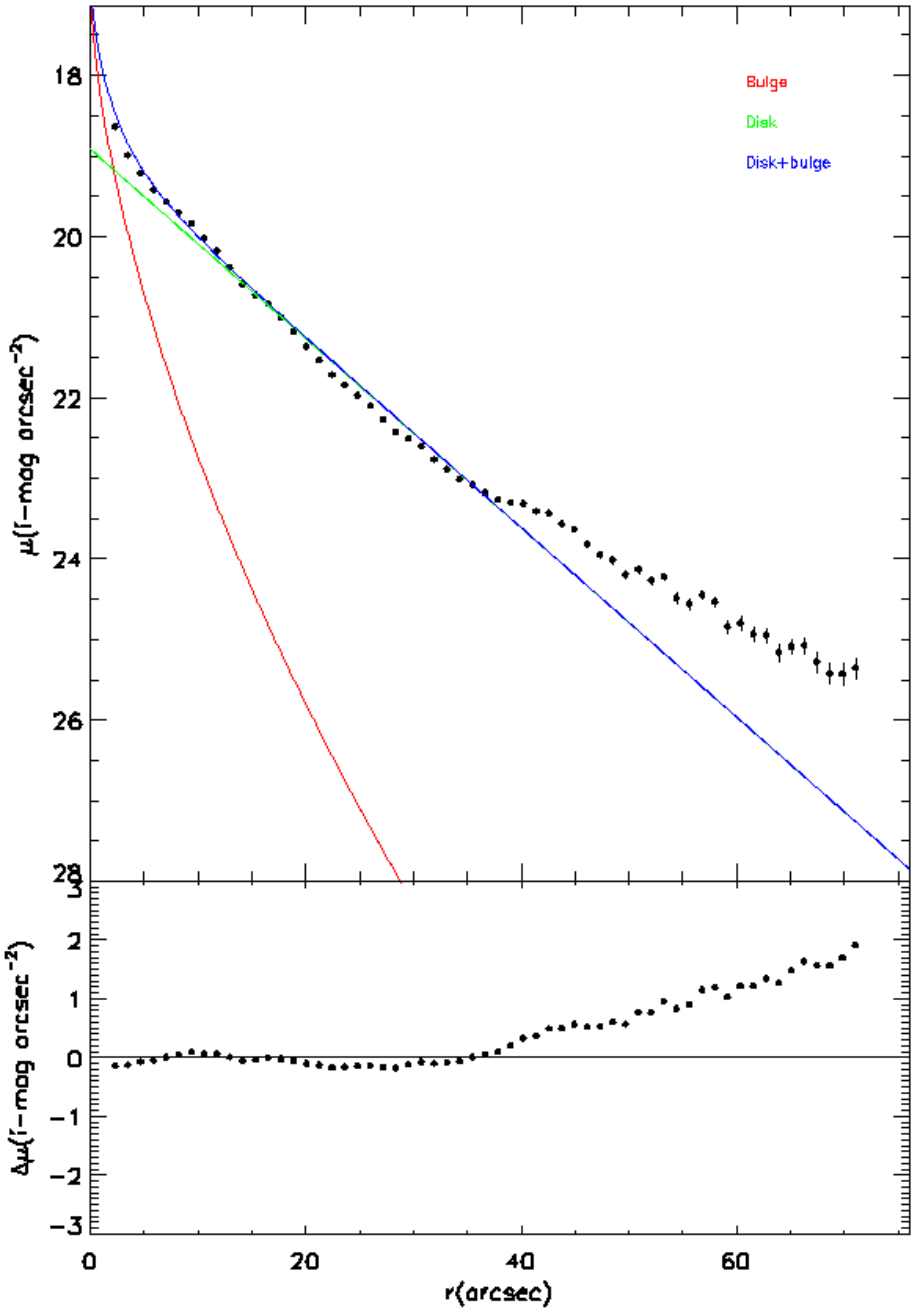


Figure 8: Luminosity profiles for galaxy CIG 33 (anti-truncation)

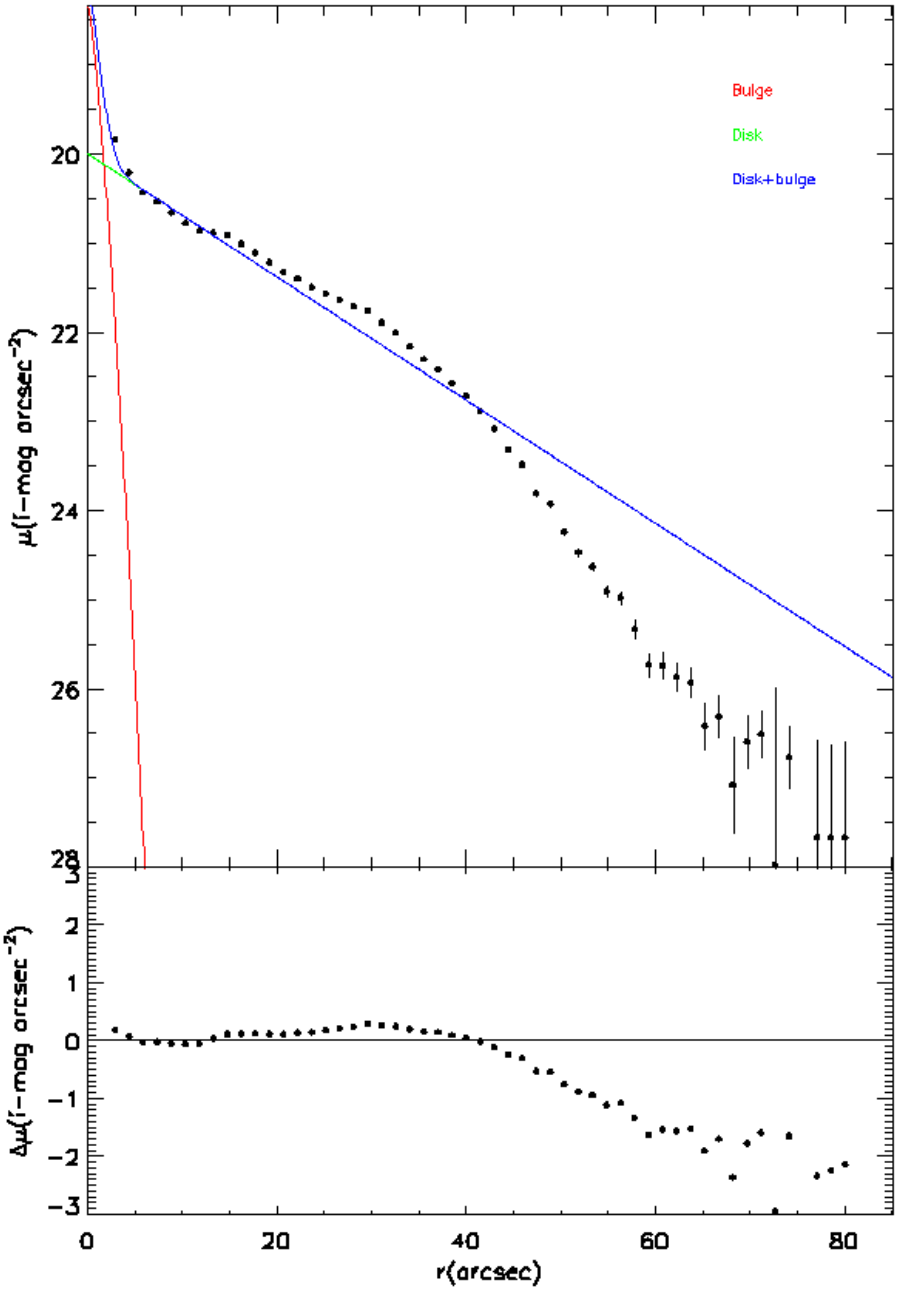


Figure 9: Luminosity profiles for galaxy CIG 281 (truncation)

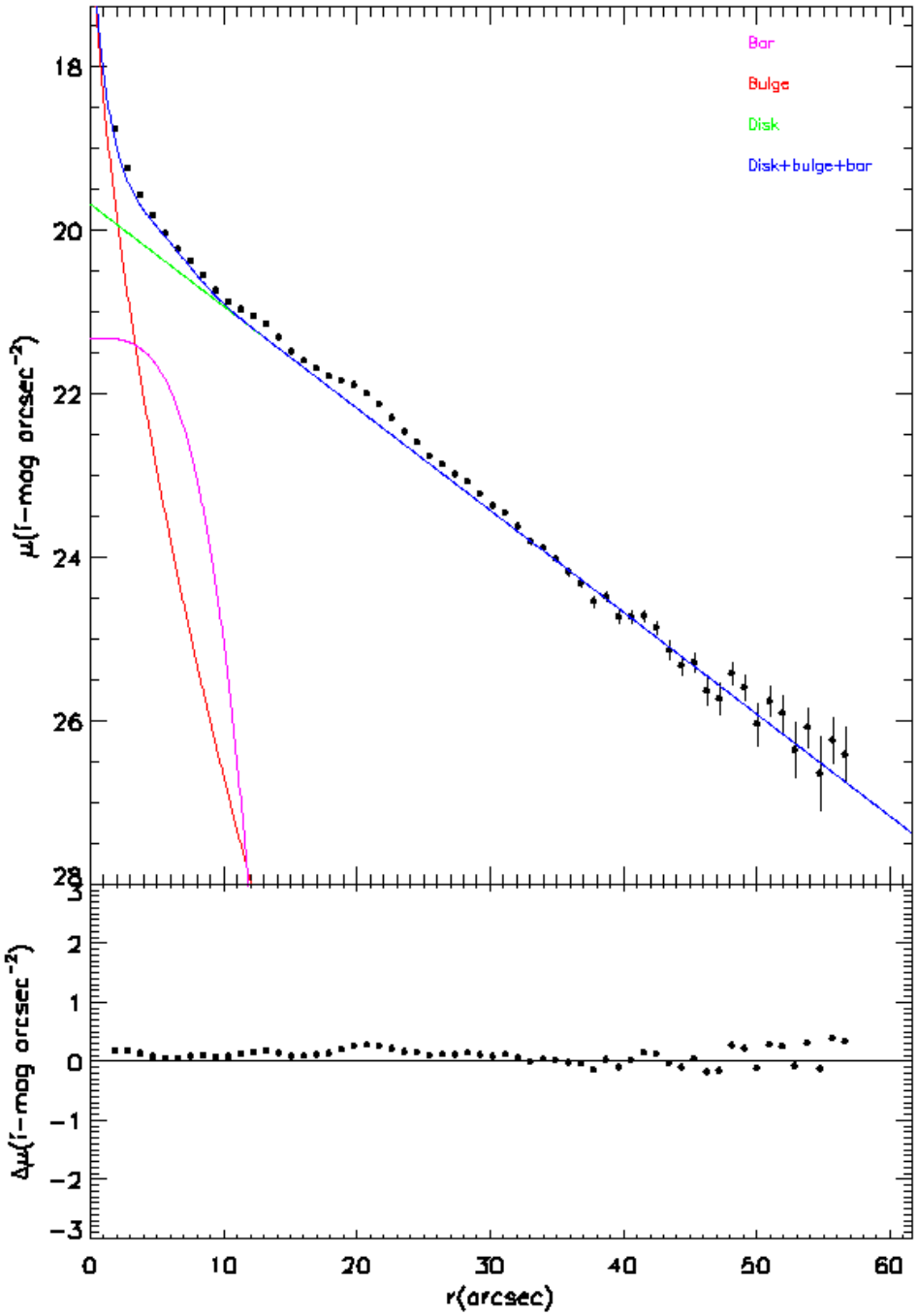


Figure 10: Luminosity profiles for galaxy CIG 520 (normal)