Handling data in large galaxy surveys for optimizing science

Elena Zucca (INAF-OABo)

Where we come from
Where we are
Where we are going

What we need
Where we come from...
First CfA Strip

$28.5 \leq \delta < 32.5$

$m_3 \leq 15.5$
90s

Scans of photographic plates - Multifiber spectroscopy

ESP survey (PI Vettolani)

CfA stops here
The ESO Slice Project (ESP) galaxy redshift survey*

I. Description and first results

G. Vettolani\textsuperscript{1}, E. Zucca\textsuperscript{2,1}, G. Zamorani\textsuperscript{2,1}, A. Cappi\textsuperscript{2}, R. Merighi\textsuperscript{2}, M. Mignoli\textsuperscript{2}, G.M. Stirpe\textsuperscript{2}, H. MacGillivray\textsuperscript{3}, C. Collins\textsuperscript{4}, C. Balkowski\textsuperscript{5}, V. Cayatte\textsuperscript{5}, S. Maurogordato\textsuperscript{2}, D. Proust\textsuperscript{5}, G. Chincarini\textsuperscript{6,7}, L. Guzzo\textsuperscript{8}, D. Maccagni\textsuperscript{8}, R. Scaramella\textsuperscript{9}, A. Blanchard\textsuperscript{10}, and M. Ramella\textsuperscript{11}

1 Istituto di Radioastronomia del CNR, via Gobetti 101, I-40129 Bologna, Italy
2 Osservatorio Astronomico di Bologna, via Zamboni 33, I-40126 Bologna, Italy
3 Royal Observatory Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK
4 School of EEEP, Liverpool John-Moores University, Byrom Street, Liverpool L3 3AF, UK
5 Observatoire de Paris, DAEC, Unité associée au CNRS, D0173 et à l’Université Paris 7, 5 Place J.Janssen, F-92195 Meudon, France
6 Osservatorio Astronomico di Brera, via Bianchi 46, I-22055 Merate (LC), Italy
7 Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy
8 Istituto di Fisica Cosmica e Tecnologie Relative, via Bassini 15, I-20133 Milano, Italy
9 Osservatorio Astronomico di Roma, via Osservatorio 2, I-00040 Monte Porzio Catone (RM), Italy
10 Université L. Pasteur, Observatoire Astronomique, 11 rue de l’Université, F-67000 Strasbourg, France
11 Osservatorio Astronomico di Trieste, via Tiepolo 11, I-34131 Trieste, Italy

Received 3 December 1996 / Accented 3 April 1997

= Bologna people

Giornate di Osservatorio - 19/02/2016
And then...

Large field CCDs

Giornate di Osservatorio - 19/02/2016
And then... Multislit spectroscopy

Giornate di Osservatorio - 19/02/2016

VIMOS@VLT
VIMOS at the ESO VLT measures the distance of 1001 distant galaxies in one single observation 28/09/2002.
The VIMOS VLT Deep Survey final data release: a spectroscopic sample of 35,016 galaxies and AGN out to $z \sim 6.7$ selected with $17.5 \leq i_{AB} \leq 24.75$

O. Le Fèvre$^{1,*}$, P. Cassata$^1$, O. Cucciati$^2$, B. Garilli$^3$, O. Ilbert$^1$, V. Le Brun$^1$, D. Maccagni$^3$, C. Moreau$^1$, M. Scodeggo$^3$, L. Tresse$^1$, G. Zamorani$^2$, C. Adami$^1$, S. Arnouts$^1$, S. Bardelli$^2$, M. Bolzonella$^2$, M. Bondi$^5$, A. Bongiorno$^6$, D. Bottini$^3$, A. Cappi$^2$, S. Charlot$^7$, P. Ciliegi$^2$, T. Contini$^8$, S. de la Torre$^9$, S. Foucaud$^{10}$, P. Franzetti$^3$, I. Gavignaud$^{11}$, L. Guzzo$^{12}$, A. Iovino$^{12}$, B. Lemaux$^1$, C. López-Sanjuan$^{1,18}$, H. J. McCracken$^7$, B. Marano$^6$, C. Marinoni$^{13}$, A. Mazure$^{†1}$, Y. Mellier$^7$, R. Merighi$^2$, P. Merluzzi$^5$, S. Paltani$^{14,15}$, R. Pellò$^8$, A. Pollo$^{16,17}$, L. Pozzetti$^2$, R. Scaramella$^4$, L. Tasca$^1$, D. Vergani$^2$, G. Vettolani$^5$, A. Zanichelli$^5$, and E. Zucca$^2$

$^*$ = Bologna people

Giornate di Osservatorio - 19/02/2016
THE \textit{zCOSMOS} 10k-BRIGHT SPECTROSCOPIC SAMPLE*

\textsc{Simon J. Lilly$^1$, Vincent Le Brun$^2$, Christian Maier$^1$, Vincenzo Mainieri$^3$, Marco Mignoli$^4$, Marco Scodeggio$^5$, Gianni Zamorani$^4$, Marcella Carollo$^1$, Thierry Contini$^6$, Jean-Paul Kneib$^2$, Olivier Le Fèvre$^2$, Alvio Renzini$^7$, Sandro Bardelli$^4$, Micol Bolzonella$^4$, Angela Bongiorno$^8$, Karina Caputi$^1$, Graziano Coppa$^4$, Olga Cucciati$^9$, Sylvain de la Torre$^2$, Loic de Ravel$^2$, Paolo Franzetti$^5$, Bianca Garilli$^5$, Angela Iovino$^9$, Pawel Kampczyk$^1$, Katarina Kovac$^1$, Christian Knobel$^1$, Fabrice Lamareille$^6$, Jean-Francois Le Borgne$^6$, Roser Pello$^6$, Yingjie Peng$^1$, Enrique Pérez-Montero$^6$, Elena Ricciardelli$^7$, John D. Silverman$^1$, Masayuki Tanaka$^3$, Lidia Tasca$^2$, Laurence Tresse$^2$, Daniela Vergani$^4$, Elena Zucca$^4$, Olivier Ilbert$^{10}$, Mara Salvato$^{11}$, Pascal Oesch$^1$, Umi Abbas$^2$, Dario Bottini$^5$, Peter Capak$^{11}$, Alberto Cappi$^4$, Paolo Cassata$^2$, Andrea Cimatti$^{11}$, Martin Elvis$^{13}$, Marco Fumana$^5$, Luigi Guzzo$^9$, Gunther Hasinger$^8$, Anton Koekemoer$^{12}$, Alexei Leauthaud$^{15}$, Dario Maccagni$^5$, Christian Marinoni$^{16}$, Henry McCracken$^{16}$, Pierdomenico Memeo$^5$, Baptiste Meneux$^8$, Cristiano Porciani$^1$, Lucia Pozzetti$^4$, David Sanders$^{10}$, Roberto Scaramella$^{18}$, Claudia Scarlata$^{11}$, Nick Scoville$^{11}$, Patrick Shopbell$^{11}$, and Yoshiaki Taniguchi$^{19}$

\textcolor{red}{=} Bologna people

\par

Giornate di Osservatorio - 19/02/2016
...even more important

Development of a significant expertise in the field

Well established group inside the Bologna Observatory

Large national and international collaborations

Funded with various INAF PRIN
(all finished)
Where we are...
Density of spectra (number per sq. deg.)

Area (sq. deg.)

Giornate di Osservatorio - 19/02/2016
2<z<6.5: probing a major epoch in galaxy assembly

What fuels star formation?
- processes to transform gas into stars
- modulated by feedback, environment (feedback)

Effect of the environment?

What contributes to the mass increase?
- evolution of the mass in stars

MASS GROWTH

Accretion

Merging

Giornate di Osservatorio - 19/02/2016
Missing: large samples of galaxies in large volumes with $2 < z_{\text{spec}} < 6.5$

Need large spectroscopic samples in large volumes

Giornate di Osservatorio - 19/02/2016
Missing: large samples of galaxies in large volumes with $2 < z_{\text{spec}} < 6.5$

At $z>2$ most studies use photometric samples

The census of galaxies so far relies on small fields

*Need large spectroscopic samples in large volumes*

Giornate di Osservatorio - 19/02/2016
VUDS: spectroscopic survey of the first phases of galaxies assembly

ESO Large Program, PI: Olivier Le Fèvre
640h allocated (~80 nights, clear)

~10000 spectra to map the Universe 10-13 Gyr ago

Understanding early galaxy assembly:
- 10,000 galaxies observed
- 14h exp.time, 3600-9300Å
- 1 deg² in 3 fields: COSMOS, ECDFS, VVDS2h
- Smart selection: photo-z and SED
- Largest spectroscopic survey in 2<z<6+

<table>
<thead>
<tr>
<th>FIELD</th>
<th>VIMOS pointings</th>
<th>Area arcmin²</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMOS</td>
<td>8</td>
<td>1800</td>
</tr>
<tr>
<td>ECDFS</td>
<td>2+1</td>
<td>675</td>
</tr>
<tr>
<td>VVDS-02</td>
<td>5</td>
<td>1125</td>
</tr>
<tr>
<td>TOTAL</td>
<td>15+1</td>
<td>3600</td>
</tr>
</tbody>
</table>

Giornate di Osservatorio - 19/02/2016
The VIMOS Ultra-Deep Survey: \(\sim 10,000\) galaxies with spectroscopic redshifts to study galaxy assembly at early epochs \(2 < z \simeq 6\)\,**

O. Le Fèvre\(^1\), L. A. M. Tasca\(^1\), P. Cassata\(^1\), B. Garilli\(^3\), V. Le Brun\(^1\), D. Maccagni\(^3\), L. Pentericci\(^4\), R. Thomas\(^1\), E. Vanzella\(^2\), G. Zamorani\(^2\), E. Zucca\(^2\), R. Amorin\(^4\), S. Bardelli\(^2\), P. Capak\(^12\), L. Cassarà\(^3\), M. Castellano\(^4\), A. Cimatti\(^5\), J. G. Cuby\(^1\), O. Cucciati\(^5,2\), S. de la Torre\(^1\), A. Durkalec\(^1\), A. Fontana\(^4\), M. Giavalisco\(^13\), A. Grazian\(^4\), N. P. Hathi\(^1\), O. Ilbert\(^1\), B. C. Lemaux\(^1\), C. Moreau\(^1\), S. Paltani\(^9\), B. Ribeiro\(^1\), M. Salvato\(^14\), D. Schaerer\(^10,8\), M. Scodéggi\(^3\), V. Sommariva\(^5,4\), M. Talia\(^5\), Y. Taniguchi\(^15\), L. Tresse\(^1\), D. Vergani\(^6,2\), P. W. Wang\(^1\), S. Charlot\(^7\), T. Contini\(^8\), S. Fotopoulou\(^9\), C. López-Sanjuan\(^11\), Y. Mellier\(^7\), and N. Scoville\(^12\)

\(^{**}\) = Bologna people (+Stephane de Barros)

Giornate di Osservatorio - 19/02/2016
What do we get?

- **80% success rate down to \( i_{AB} = 25 \)**
- Absorption as well as emission spectra
- Interesting outliers
  - Contamination along the LOS
- SED fitting using em. line templates
  - Exceptionnal set of multi-\( \lambda \) data (HST, Subaru, UltraVista, Spitzer…)
  - SFR, \( M_{\text{star}} \), E(B-V), Age,…
- SED+spectra fitting
- Redshift distribution as expected
What do we get?

- 80% success rate down to $i_{AB}=25$
- Absorption as well as emission spectra
- Interesting outliers
  - Contamination along the LOS
  - Exceptionnal set of multi-$\lambda$ data (HST, Subaru, UltraVista, Spitzer...)
  - SFR, $M_{\text{star}}$, E(B-V), Age,...
- SED+spectra fitting
- Redshift distribution as expected
Absorption & emission line galaxies to $z \sim 6$

Individual spectra $i_{AB} \leq 25$, a very faint sample

Giornate di Osservatorio - 19/02/2016
What do we get?

- 80% success rate down to $i_{AB} = 25$
- Absorption as well as emission spectra
- **Interesting outliers**
  - Contamination along the LOS
- SED fitting using em. line templates
  - Exceptionnal set of multi-λ data (HST, Subaru, UltraVista, Spitzer…)
  - SFR, $M_{star}$, E(B-V), Age,…
- SED+spectra fitting
- Redshift distribution as expected

Superimposition on the line of sight

Giornate di Osservatorio - 19/02/2016
What do we get?

- 80% success rate down to $i_{AB} = 25$
- Absorption as well as emission spectra
- Interesting outliers
  - Contamination along the LOS
- SED fitting using em. line templates
  - Exceptionnal set of multi-λ data (HST, Subaru, UltraVista, Spitzer...)
  - SFR, $M_{\text{star}}$, $E(B-V)$, Age,...
- SED+spectra fitting
- Redshift distribution as expected

R. Thomas et al.
What do we get?

- 80% success rate down to $i_{AB}=25$
- Absorption as well as emission spectra
- Interesting outliers
  - Contamination along the LOS
- SED fitting using em. line templates
  - Exceptional set of multi-$\lambda$ data (HST, Subaru, UltraVista, Spitzer...)
  - SFR, $M_{\text{star}}$, $E(B-V)$, Age,…
- SED+spectra fitting
- Redshift distribution as expected
Before VUDS

All spec. surveys before VUDS

- VVDS-Deep
- Steidel + BM/BX+LBG
- VLRS
- GOODS@VLT
- zCOSMOS-Deep
- Ouchi+10 LAE z~6.5
- Stark+10 dropout

Redshift vs. N\textsubscript{Galaxies}

Giornate di Osservatorio - 19/02/2016
All spec. surveys with VUDS

Redshift

N_Galaxies

VUDS
VVDS-Deep
Steidel+ BM/BX+LBG
VLRS
GOODS@VLT
zCOSMOS-Deep
Ouchi+10 LAE z~6.5
Stark+10 dropout

Giornate di Osservatorio - 19/02/2016
VUDS \sim 7500 spectra of galaxies at $z > 2$: \sim 3$Gyr of evolution in one glance
VUDS observations completed in 2013
reductions completed in 2014
up to now ~20 refereed papers
(10 published, 8 in press/submitted)
many more in preparation

(some results in Sandro's talk)

Mobility funded with INAF PICS 2014-2015-2016
(Italy – France agreements)
Where we are going...
PI Y. Mellier
~1200 people involved

Wide Survey ~15,000 deg²
Deep Survey ~40 deg²

~2 × 10⁹ objects

Main responsibilities

OU-LE3
Leader of Internal Data Validation WP

SWG Galaxy evolution
Coordinator of Distribution functions WP

+ various roles in other WPs
4 THE EUCLID GROUND SEGMENT

4.1 Overview

Euclid is expected to return roughly four times as much data as Gaia. The EMC in collaboration with ESA provides a framework for the end-to-end handling of this vast amount of data. In this section a broad overview is given of the Euclid ground segment and its main elements: MOC, SOC, and the EMC provided SDCs.

The Euclid data processing system is organised in sequential processing steps of increasing sophistication. With each step is associated a data processing level or “data level”. Data levels consist of all data produced by the corresponding data processing step including intermediate data. Each data level has corresponding quality controls. The list of data levels includes:

- Level 1 data. Unpacked and checked telemetry and housekeeping data
- Level 2 data. Calibrated data and intermediate data products produced during the calibrations. Calibrated data have all instrumental fingerprints removed. Dithered images have been co-added; spectra have been extracted
- Level 3 data. Science-ready data products, mostly catalogues, which fulfill the science objectives of the mission. Level 3 includes pre-defined science data products (3D galaxy power spectra, dark matter power spectra, tomography, high order statistics, mass reconstruction maps...) but does not include data analysis beyond the production of catalogues and basic 2-point statistics or cosmological interpretation of data (joint analyses of data, dark energy studies, cosmological parameters, growth and growth rate of structures, galaxy biasing, test gravity, neutrino mass constraints, galaxy clustering, etc...).
- External Data. Quality-controlled data from existing missions and ground-based surveys which are used for calibrations, photometric redshift derivations, and simulation validations before and during operations.
- Level S (Simulation) data. Pre-launch simulations and modeling impacting on calibrations and observing strategies. Massive Monte Carlo simulations are likely to be required post launch to assess systematic effects and derive meaningful uncertainties on the final cosmological parameters.
- Level Q (Quick-release) data which the SOC generates from the raw telemetry and uses to perform a first level quality check. Level Q data represent products suitable for most purposes in Astronomy, except for the core Cosmology objectives of the mission. They will be distributed to the general scientific community during the proprietary period (see section 5). Level Q data are superseded by Level 1-3 data once these are released to the general scientific community.
4 THE EUCLID GROUND SEGMENT

4.1 Overview
Euclid is expected to return roughly four times as much data as Gaia. The EMC in collaboration with ESA provides a framework for the end-to-end handling of this vast amount of data. In this section a broad overview is given of the Euclid ground segment and its main elements: MOC, SOC, and the EMC provided SDCs.

The Euclid data processing system is organised in sequential processing steps of increasing sophistication. With each step is associated a data processing level or “data level”. Data levels consist of all data produced by the corresponding data processing step including intermediate data. Each data level has corresponding quality controls. The list of data levels includes:

- Level 1 data. Unpacked and checked telemetry and housekeeping data
- Level 2 data. Calibrated data and intermediate data products produced during the calibrations. Calibrated data have all instrumental fingerprints removed. Dithered images have been co-added; spectra have been extracted

- Level 3 data. Science-ready data products, mostly catalogues, which fulfil the science objectives of the mission. Level 3 includes pre-defined science data products (3D galaxy power spectra, dark matter power spectra, tomography, high order statistics, mass reconstruction maps...) but does not include data analysis beyond the production of catalogues and basic 2-point statistics or cosmological interpretation of data (joint analyses of data, dark energy studies, cosmological parameters, growth and growth rate of structures, galaxy biasing, test gravity, neutrino mass constraints, galaxy clustering, etc...).

For most purposes in Astronomy, except for the core Cosmology objectives of the mission. They will be distributed to the general scientific community during the proprietary period (see section 5). Level Q data are superseded by Level 1-3 data once these are released to the general scientific community.
**Euclid SGS LE3**

**Processing Function Requirements Specification Document**

VMSP-ID: The spectroscopic redshift selection function (Spectroscopic Visibility Mask)

SEL-ID: The estimate of the purity and completeness of galaxy samples and the selection of a catalog that maximises both.
Internal Data: processing functions for deriving
- Visibility masks
- Selection function
- Luminosity and Mass function

We must
- Choose the best algorithms (requirements from SWGs)
- Code the algorithms in specific languages
- Follow the migration to Science Data Centers

Implementation and Validation procedure
Our role changed during the years...
What we need...
Need of specialized man power (software engineers)

No man power assigned

No specific budget assigned

Strict deadlines
Need of specialized man power (software engineers)

No man power assigned

No specific budget assigned

Strict deadlines

Euclid science will be based on OU-LE3 outputs